

AN ASSESSMENT PROCEDURE FOR DETECTING  
GIFTEDNESS USING AVAILABLE DATA

BY

RANDY SCHNELL

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL  
OF THE UNIVERSITY OF FLORIDA IN  
PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1987

## ACKNOWLEDGEMENTS

I would like to thank Larry Loesch, Ph.D., for his guidance and assistance as chairperson of my doctoral committee. Dr. Loesch graciously assumed the role of chairperson during a period when my study was in disarray and not progressing satisfactorily. I also wish to thank Robert Jester, Ph.D., whose consultation assisted in development of the framework of this study, and Janet Larsen, Ed.D., who has supported my doctoral work since 1982 when she agreed to be my advisor. I would also like to acknowledge Linda Crocker, Ph.D., whose patience, diligence, and expertise in measurement were vital to the study and greatly appreciated.

A number of other people who contributed to this study also deserve recognition. These include John Hilderbrand, Ph.D., Hugh Morehouse, Robert Haines, Ed.D., Grace Hutchinson, Michael Selby, Lois Rudloff, and Denise Landau, Ph.D.

I wish to express special thanks to my parents and other family members for their support.

## TABLE OF CONTENTS

	<u>Page</u>
ACKNOWLEDGEMENTS . . . . .	ii
ABSTRACT . . . . .	v
 CHAPTER	
I      INTRODUCTION . . . . .	1
Problem . . . . .	3
Giftedness . . . . .	4
Purpose . . . . .	5
Research Questions . . . . .	5
Definition of Terms . . . . .	6
Theoretical Rationale . . . . .	9
Need for This Study . . . . .	15
Overview of the Remainder of the Study . . . . .	17
II     LITERATURE REVIEW . . . . .	18
Support for the Problem . . . . .	18
Group IQ Tests . . . . .	19
Intelligence Quotient Short Forms . . . . .	21
Intelligence Quotient Screening Tests . . . . .	25
Achievement Tests . . . . .	30
Summary . . . . .	39
Instruments Used in Study . . . . .	39
Wechsler Intelligence Scale for Children-Revised (WISC-R) . . . . .	39
Slosson Intelligence Test SIT) . . . . .	41
Comprehensive Test of Basic Skills (CTBS)/Test of Cognitive Skills (TCS) . . . . .	44
Stanford-Binet Intelligence Scale (S-B) . . . . .	46

III	METHODOLOGY . . . . .	49
	Overview . . . . .	49
	Population and Sample . . . . .	49
	Assessment Procedures . . . . .	51
	Research Procedures . . . . .	53
	Data Analysis . . . . .	55
	Methodological Limitations . . . . .	58
IV	RESULTS . . . . .	62
	Phase I--Item Selection . . . . .	62
	Phi Coefficients . . . . .	62
	Index of Discrimination . . . . .	66
	Cutoff Scores . . . . .	68
	Phase II--Cross Validation . . . . .	70
	Application of Cutoff Scores . . . . .	70
	Kappa Comparisons . . . . .	74
	Summary of Results . . . . .	78
	Summary of Results . . . . .	78
V	DISCUSSION . . . . .	80
	Research Questions . . . . .	80
	Phase I . . . . .	81
	Common Items . . . . .	81
	IRT Parameters . . . . .	82
	Phase II . . . . .	83
	Conclusions, Implications, and Limitations . . . . .	85
	Sampling . . . . .	85
	Generalizability . . . . .	87
	Screening Accuracy . . . . .	88
	Item Validity . . . . .	88
	Cutoffs . . . . .	90
	Summation . . . . .	91
	REFERENCES . . . . .	92
	BIOGRAPHICAL SKETCH . . . . .	104

Abstract of Dissertation Presented to the Graduate School  
of the University of Florida in Partial Fulfillment of  
the Requirements for the Degree of Doctor of Philosophy

AN ASSESSMENT PROCEDURE FOR DETECTING  
GIFTEDNESS USING AVAILABLE DATA

By

Randy Schnell

December 1987

Chairman: Larry Loesch, Ph.D.

Major Department: Counselor Education

The purpose of this study was to examine a new psychometric screening procedure designed to discriminate intellectually "gifted" seventh graders from other high achieving seventh graders. All 179 students in the research sample had been assessed to determine eligibility in the Hillsborough County Public Schools "gifted" program. An accurate and efficient screening procedure was necessary in order to delete from time-consuming, formal assessment those students who were unlikely to meet intellectual criteria for gifted program placement. The Slosson Intelligence Test (SIT), previously used to screen students referred to the gifted classes, had proved inefficient for this purpose. A plethora of research on

methods used to identify gifted students revealed that most methods were less than accurate.

The Comprehensive Test of Basic Skills and Test of Cognitive Skills were item analyzed on a sample of 179 seventh grade subjects using two correlational procedures. Results from the item analyses yielded 46 items from the phi analysis and 14 from the index of discrimination analysis that discriminated gifted from not gifted students, as classified by the WISC-R or Stanford Binet. Cross validation was conducted on a second sample of 61 students to determine (a) the accuracy of the new screening procedure (NP) compared to the SIT and (b) at what cutoff points either the NP or SIT was more accurate. In cross validation, total scores of both item subsets were compared with SIT. The items obtained from the index of discrimination analyses were found to be the best predictors of "gifted" intelligence for the sample. There was some ambiguity regarding an optimal cutoff score; however, a score .33 standard deviations below the mean score was found to generally yield most accurate predictions.

## CHAPTER I INTRODUCTION

The tremendous increase since 1970 in school programs for the intellectually gifted has led most states to establish guidelines or requirements for program eligibility (Karnes & Brown, 1979; Kolloff & Feldhusen, 1984). One common requirement is for individual intellectual evaluation of gifted program candidates (Chambers, Barron, & Sprecher, 1980; Sternberg, 1982; Vernon, Adamson, & Vernon, 1977). Yarborough and Johnson (1983) reported the use of individual intelligence test minimum scores as an eligibility requirement in at least 73% of the nation's gifted student programs. Although educators may recommend and actually believe in the importance of a broader view of giftedness, this more narrow definition (i.e., emphasizing IQ) is frequently employed because of pragmatic concerns (Jenkins-Freidman, 1982). The "intelligence quotient" has therefore emerged as the primary criterion for gifted program eligibility (Barklay, Phillips, & Jones, 1983; Birch, 1984; Guilford, 1975; Karnes, Edwards, & McCallum, 1986).

The individual intelligence testing requirement, coupled with federal regulations mandating services for

children with special educational needs, has placed a burden on local school districts to identify gifted students through accurate referral procedures and efficient use of testing personnel (Karnes & Brown, 1979). As a result, school systems are currently faced with demands for intellectual assessment of children which are often greater than can be met by the qualified examiners available (Crofoot & Bennett, 1980; Fell & Fell, 1982). There is a demand for screening procedures designed to facilitate the referral process by maximizing the use of individual testing time while minimizing errors in identification (Jenkins-Friedman, 1982; Kramer, Markley, Shanks, & Ryabik, 1983; Rust & Lose, 1980; Stephens & Gibson, 1963).

The Wechsler Intelligence Scale for Children--Revised (WISC-R) and Stanford-Binet Intelligence Scale (S-B) are the most widely used individual tests of children's intelligence (Bryan & Bryan, 1975; Salvia & Ysseldyke, 1978; Wikoff, 1978). Screening tests which estimate or predict intelligence scores on the WISC-R and Stanford-Binet have been studied for use among gifted school populations since the 1950s (Pegnato & Birch, 1959; Sheldon & Manolakes, 1954).



### Problem

The research problem addressed concerned the accurate identification of intellectually gifted students from a screening procedure prior to administration of an individual intelligence test. The expense of administering individual intelligence tests has necessitated the screening of potentially gifted students in an effort to delete from formal testing those who probably do not possess gifted intelligence (Rust & Lose, 1980; Stenson, 1982). Inaccurate screening has resulted in not-gifted students receiving the time-consuming tests and in some gifted students being excluded from testing.

Screening procedures such as the Slosson Intelligence Test (Dirkes, Wessels, Quaforth, & Quenon, 1980; Grossman & Johnson, 1983; Karnes & Brown, 1979; Rust & Lose, 1980), the Ammons Quick Test (DeFilippis & Fulmar, 1980; Hirsch & Hirsch, 1980), short forms of the WISC-R (Bersoff, 1971; Elman, Blixt, & Sawicki, 1981; Karnes & Brown, 1981; Killian & Hughes, 1978; Kramer et al., 1983), Guilford's Structure of the Intellect (Pearce, 1983), the Peabody Picture Vocabulary Test (Mize, Smith, & Callaway, 1979; Wright, 1983), and group IQ tests (Blosser, 1963; Grossman & Johnson, 1983; Pagnato & Brich, 1959; Sheldon & Manolakes, 1954) have all been shown to be more or less inaccurate and/or inefficient for screening high ability

students. Chambers (1960) and Schena (1963) reported somewhat more encouraging results with academic skill measures as predictors of gifted intelligence. A more accurate and efficient means for screening gifted students needs to be found.

### Giftedness

The characteristics associated with gifted intelligence are almost as numerous as the students themselves (Tuttle & Becker, 1980). Qualitative trait differences between gifted and not-gifted children are indicated frequently in the literature (Barrington, 1979; Dirkes, 1981; Gensley, 1975; Male & Perrone, 1979; Ricca, 1984; Ryan, 1982; Sternberg, 1982); however, essential to a discussion of intellectual giftedness in children is their classification according to an IQ test cutoff score. Gifted classifications by IQ cutoff scores are employed in an attempt to objectify classification criteria and school placement decisions. Classification by IQ score may misleadingly imply that qualitative differences between gifted and not-gifted children are necessarily demarcated by an artificial cutoff (Braden, 1985). In this study cutoff scores are used to quantify gifted intelligence according to educational criteria and not to define learning style, motivation, or other personality traits associated with "giftedness."

### Purpose

The primary purpose of this study was to determine whether a large, group-administered achievement/ability test battery possesses items that, in combination, yield a score that accurately predicts gifted classification as measured by the WISC-R or Stanford-Binet. A secondary purpose of this research was to determine if the new screening procedure (NP) classifies gifted and not-gifted seventh graders more accurately than the Slosson Intelligence Test (SIT) and at what cutoffs these classifications are most accurate.

The new screening procedure developed for this study consisted of a subset of items selected from norm referenced, group-administered tests of academic aptitude and achievement. The aptitude test used in this study was the Test of Cognitive Skills (TCS) and the achievement battery was the Comprehensive Test of Basic Skills (CTBS). The TCS and CTBS have been normal on the same sample and are typically administered in concurrent testing sessions.

### Research Questions

The following research questions were addressed:

1. Can an accurate predictor of gifted IQ classification on the WISC-R/S-B be derived from an instrument

composed of items on the CTBS and TCS in a situation in which giftedness is viewed as a dichotomous variable?

2. Is the NP more accurate than the SIT in classifying gifted and not-gifted seventh graders?

3. At what cutoff point(s) is the NP more accurate than the SIT most accurate in classifying gifted and not-gifted seventh graders?

### Definition of Terms

Comprehensive Test of Basic Skills (CTBS). The CTBS is a "series of norm-referenced, objective-based tests for kindergarten through twelfth grade. The series is designed to measure achievement in the basic skills commonly found in state and district curricula" (CTB/McGraw-Hill, 1984, p. 1). At the junior high school levels the content areas are reading, spelling, language, mathematics, reference skills, science, and social studies.

Full Scale IQ (FS-IQ). Full Scale IQ is the derived intelligence quotient on the Wechsler Intelligence Scale for Children--Revised and on the Stanford-Binet Intelligence Scale.

General intelligence. General intelligence is a set of general cognitive operations measured as the overall ability required for success on IQ tests. General

intelligence is comprised of those traits commonly measured by IQ test items.

Gifted intelligence. Gifted intelligence is measured as an intelligence quotient that falls in the Very Superior category or two standard deviations above the test mean. For this research, gifted IQ=130 -1 SEM (where SEM=3 IQ points) on the WISC-R, and IQ=132 -1 SEM (where SEM=5 IQ points) on the Stanford-Binet, or IQ=127. (This definition, which includes a SEM, is based on guidelines from Hillsborough County Florida School District).

High-achievers. Students who, because of superior academic performance, have been referred for gifted program testing, but who are assessed as not intellectually gifted are referred to as high-achievers.

Intelligence. Intelligence is "the aggregate global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment" (Wechsler, 1958, p. 7). Intelligence is comprised of intellectual factors such as "abstract reasoning, verbal, spatial, numerical, and other factors" (Wechsler, 1950, p. 78), and nonintellectual factors consisting of "capacities and bents dependent upon temperament and personality which are . . . factors of personality itself" (Wechsler, 1950, p. 78).

Intelligence quotient. An intelligence quotient is a derived total or Full Scale score on an intelligence test.

Performance IQ (P-IQ). Performance IQ is a subscale IQ score on the Wechsler Intelligence Scale for Children--Revised that represents the examinee's nonverbal reasoning or perceptual organization (Kaufman, 1975).

Reliability. Reliability is the squared population correlation between the individual's obtained score and the individual's hypothetical true score. Reliability is "the proportion of true-score variance in scores on a particular test at the time it was taken" (Jensen, 1980, p. 260).

Slosson Intelligence Test (SIT). The SIT is an individually administered intelligence test which requires little specialized training to administer, only about 20 minutes to administer and score, and yields IQ scores "which are close approximations to the Stanford-Binet IQ [scores]" (Slosson & Jensen, 1982, p. 1).

Stanford-Binet Intelligence Scale (S-B). The S-B is one of the two individually administered intelligence tests used in this study to measure gifted intelligence. (See page 41 for a detailed description.)

Test of Cognitive Skills (TCS). The TCS is "an ability test designed to assess a student's academic aptitude and thereby predict the student's level of success in school. Emphasis in TCS is placed on" . . . problem solving, discovering relationships, evaluating, and

remembering" (CTB/McGraw-Hill, 1984, p. 1). The TCS is administered concurrently with the CTBS.

Validity. Validity refers to the appropriateness of inferences from test scores or other forms of assessment. Validity deals with how faithfully the scores represent a domain of skill, knowledge, or of a trait being measured (American Psychological Association, 1985).

Verbal IQ (V-IQ). Verbal IQ is a subscale IQ score on the Wechsler Intelligence Scale for Children--Revised that represents the examinee's verbal reasoning or verbal comprehension (Kaufman, 1975).

Wechsler Intelligence Scale for Children--Revised (WISC-R). The WISC-R is the individually administered IQ test predominantly used in this study to measure gifted intelligence. (See page 35 for a detailed description.)

### Theoretical Rationale

Salvia and Ysseldyke (1978) have pointed out that there is a hypothetical domain of items that may be used to assess intelligence and that items may be drawn from various sources. For example, WISC-R Information subtest items are drawn from a domain of achievement oriented items that measure specific content of learning acquired, in large part, through formal education. This overlapping of achievement and aptitude test item content has been



demonstrated empirically by Anastasi (1976). Her examination of the content of several current instruments classified as achievement and intelligence tests revealed similarity in their content. Supporting this finding she contended that it has long been known that IQ tests correlate about as highly with achievement tests as different IQ tests correlate with each other. Further, one of the most frequently employed means of validating IQ tests is to compare them with measures of achievement.

In another attempt to show some common elements of achievement and intelligence, Gronlund (1976) compared factors measured by both reading readiness tests and IQ tests. These elements included

1. visual discrimination--identifying similarities and differences in words or pictures;
2. auditory discrimination--identifying similarities and differences in spoken words;
3. verbal comprehension--demonstrating an understanding of the meaning of words, sentences, and directions; and
4. copying--demonstrating skills in reproducing geometric forms.

In his analysis of the WISC-R Verbal Scale, Kaufman (1979) identified item content that reflects properties of achievement tests. For example, the Information subtest



of the WISC-R measures acquired knowledge and is influenced by outside reading and school learning. The Similarities subtest was also found to be subject to reading and vocabulary knowledge. Likewise, the other Verbal Scale subtests were found to have strong components of acquired knowledge.

In one of a very few studies in which IQ test item content was related to specific academic skills, Washington, Engelmann, and Bereiter (1969) conducted an item analysis of the Stanford-Binet Intelligence Scale and attempted to construct an academic curriculum from it. After the curriculum was presented to students an achievement test was administered. Results showed that the prelearned S-B items were positively correlated with post-test achievement items for particular learning tasks. In a second phase of the study no pretest was administered. However, subsequent to the curriculum presentation and post-test administration the Stanford-Binet was given. The achievement test results were found to accurately predict S-B scores in terms of items responses. Results suggested content validity across the IQ and achievement measures.

A number of other researchers have investigated the relationship between intelligence and achievement measures and found them to be positively correlated (Hale, 1978; Hartlage & Steele, 1977; Reschley & Reschley, 1979;

Schwarting & Schwarting, 1977) or to possess significant overlap in their factor loadings (Carroll, 1966; Dean, 1977; Grossman & Johnson, 1982; Horn, 1970; Stewart & Morris, 1977; Undheim, 1976; Vernon, 1961, 1969; Wikoff, 1978).

The TCS, in terms of its title and stated purpose, encompasses a construct closely related to the WISC-R and Stanford-Binet in that all three were designed and have been found to predict academic attainment (Hurrocks, 1964; CTB/McGraw-Hill, 1984; Terman & Merrill, 1973; Wechsler, 1974). As described in its 1983 Technical Report, the TCS subtests (Sequences, Analogies, Memory, and Verbal Reasoning) load on factors consistent with those described in WISC-R and Stanford-Binet studies (Kaufman, 1979).

The TCS, according to its constructors, measures "a number of cognitive abilities included in various theories, however, like the WISC-R and Stanford-Binet, emphasis is placed on the kinds of reasoning and retention skills necessary for school success" (CTB/McGraw-Hill, 1984, p. 5). The CTBS similarly measures a variety of academic skill areas shown to be positively correlated with the TCS. The TCS and CTBS together appear to assess a theoretical construct common to the WISC-R and Stanford-Binet. Therefore, the items chosen from the CTBS and TCS

should conform to that construct, thus allowing discrimination of gifted from not-gifted students as do the WISC-R or Stanford-Binet.

Support for the common item content of the CTBS, TCS, WISC-R, and Stanford-Binet will be provided here by illustrations of actual items found on these tests. Items are categorized arbitrarily according to subtest classification and to face commonalities. A brief description and reference location will be given for those non-verbal test items that cannot be readily reproduced in this format. Items indicated are those designed for average and above average seventh graders. Test items are printed in bold-face.

1. WISC-R, Vocabulary-instructions: What does \_\_\_\_\_ mean?  
**rivalry**

CTBS, Vocabulary (instructions: Choose the word or phrase that means the same, . . . , as the underlined word.)

**their opponent**

- A. foe
- B. employee
- C. architect
- D. assistant

2. WISC-R, Arithmetic  
**Tony bought a second hand bicycle for \$28.00. He paid  $\frac{2}{3}$  of what the bicycle cost new. How much did it cost new?**

CTBS, Mathematics Concepts and Applications

**Homer's recipe will make 48 sugar cookies. He made  $\frac{3}{4}$  of this recipe for a party. How many cookies were for each of the 18 people at the party?**

## 3. Stanford-Binet, Arithmetic Reasoning

If a man's salary is \$20 a week and he spends \$14 a week, how long will it take him to save \$300?

CTBS, Mathematics Concepts and Applications

To pay for groceries, Scott, Marvin, and Carol each gave the clerk \$1.35. The clerk gave them \$.45 in change. How much did the groceries cost?

- F. \$1.80
- G. \$3.60
- H. \$4.05
- J. \$4.50

## 4. Stanford-Binet, Vocabulary

What does "Brunette" mean?

CTBS, Vocabulary (instructions: Choose the word or phrase that means the same . . . as the underlined word.)

successful merchant

- F. parade
- G. business
- H. customer
- J. shopkeeper

## 5. WISC-R Picture Arrangement (instructions: "... I want you to arrange these pictures in the right order to tell a story that makes sense.")

TCS, Sequence (instructions: "... choose the part that would continue the pattern or sequence.") Various visual stimuli are presented such as letters, numbers or geometric shapes.

## 6. WISC-R, Similarities

In what way are a telephone and a radio alike?

TCS, Verbal Reasoning (paraphrased instructions: The words in the top and bottom rows are related in the same way. Find the word that completes the bottom row of words.)

radio electricity music  
paper \_\_\_\_\_ newspaper

- F. ink
- G. story
- H. reporter
- I. typewriter

## 7. Stanford-Binet, Problems of Fact

"An Indian who had come to town for the first time in his life saw a boy riding along the street. As the boy rode by the Indian said, 'The white boy is lazy; he walks sitting down! What was the boy riding on that caused the Indian to say, 'He walks sitting down'?"

TCS, Verbal Reasoning (instructions: find the true statement)

All bicycles have gears.

Some bicycles have ten speeds.

Maria has a bicycle.

F. Maria likes her bicycle.

G. Maria's bicycle has gears.

H. Maria's bicycle goes too fast.

J. Maria's bicycle has ten speeds.

## 8. Stanford-Binet, Induction

(This is a sequential test in which paper is folded and holes cut in it by the examiner. The student must deduce a pattern to predict how many holes will result from each cut.)

TCS, Sequences

(Students are presented sequential patterns that are incomplete. The student must deduce the pattern and predict the final pattern.)

### Need for This Study

Teachers at the seventh-grade level typically experience more difficulty identifying potentially gifted students than do teachers at lower grade levels (Schnell, 1982). Consequently, junior high school teachers refer a larger proportion of not-gifted students for evaluation than do teachers of elementary school students. This phenomenon is believed to result from more limited contact between individual teachers and students and from the fact that the pool of potentially gifted students from which

junior high school teachers choose contains none of the students who have been identified as gifted during previous years. Junior high school teachers must base their referrals on "the best of the rest." Clearly the most advantageous time to identify students for a junior high school curriculum for the gifted is when they enter seventh grade.

The new procedure (NP) was developed by selecting items from popular group tests administered at seventh-grade, measuring both achievement and cognitive ability. The rationale for choosing screening items from standardized group tests is that this approach to student screening is both time- and cost-efficient. All students in the population school district take the CTBS and TCS annually and the screening data are readily available without additional tests or testing time being needed.

It was believed that of the 460 CTBS and TCS questions typically administered to seventh graders, there existed a subset of items that would accurately predict gifted IQ on the WISC-R or S-B. Because the number of CTBS/TCS items is large, sampling of a wide range of skills and abilities is possible. As the diversity of the items increases, so does the general ability measured by the total test (Kaufman, 1979).

The utility of finding a small pool of items that correlates highly with gifted intelligence is that (a) current forms of the CTBS/TCS selected items may be retained for group administration to future gifted class candidates, (b) if a gifted candidate is not present for CTBS/TCS testing the entire test will not have to be administered as a gifted screening, and (c) the procedure of analyzing a test in this way might be useful for predicting intelligence (or other traits) among other populations, utilizing these or other tests.

#### Overview of the Remainder of the Study

The subsequent content of this study is divided into four chapters. In Chapter II a review of related literature is presented. A description of the methodology used for the research comprises Chapter III. Research results are presented in Chapter IV and results are discussed in Chapter V.



## CHAPTER II LITERATURE REVIEW

### Support for the Problem

The psychometric screening of "gifted intelligence" is beset by problems of not only time- and cost-efficiency but of predictive accuracy. In the relevant literature it is suggested that these problems exist for a variety of screening methods and procedures. These problems are not of recent origin, however. As early as 1959, Pegnato and Birch found that sufficient psychological services were rarely available to test all of the gifted class candidates, thereby necessitating procedures for screening prior to formal testing. Accordingly, the authors conducted an investigation of the "relative efficiency and effectiveness" (p. 300) of seven procedures for locating gifted children in junior high schools: teacher ratings, class rank, creative ability in art or music, student council membership, superiority in mathematics, group achievement, and group IQ. Seven hundred eighty-one metropolitan school district students were selected for participation in the study on the basis of high ratings in one or more of the seven categories. All of the participants received the Stanford-Binet. Scores on this



intelligence test were used as a criterion reference for the designation of those children who were, indeed, gifted. For each of the seven screening procedures, effectiveness was judged by the percentage of gifted children located; efficiency was defined as the ratio between total number of gifted students and students predicted as gifted. Of the 781 subjects, 91 (6.5% of the school population) were judged to be gifted. These results indicated that, among the seven methods, group IQ and achievement tests were the better predictors, providing the best possible combination of effectiveness with efficiency. Other methods, such as honor role inclusion, were fairly effective, but their efficiency was poor.

Even though Pagnato and Birch (1959) were largely unsuccessful in finding an effective predictor of gifted IQ, a substantial amount of research has focused on screening the gifted since that time. The following sections of this literature review are concentrated on four procedures for gifted student identification that involve group IQ tests, IQ short forms, IQ screening tests, and achievement test scores.

#### Group IQ Tests

The largely unsuccessful attempt by Pagnato and Birch (1959) to predict gifted intelligence using group IQ tests was followed one year later by a similar study. Chambers

(1960) sought a screening instrument for use in a Michigan school district. Using the IPAT (Cattell's test of general intelligence), the California Test of Mental Maturity, the SRA Primary Abilities Test, The Kuhlman-Anderson Intelligence Test, and the WISC, Chambers tested 39 children in grades three through six. For each screening test, a cutoff was calculated above which all gifted students (WISC IQ>124) would be identified. The accuracy of each screening procedure was established at 100%, and the efficiency was then determined based on the number of not-gifted students misclassified by the screening procedure as gifted. The results revealed that the SRA test and the Kuhlman-Anderson could be ranked respectively as the most and least efficient, and that between 20% and 57% of the students predicted as gifted were not.

Three years after Chambers' study, Blosser (1963) tested 187 ninth graders on the Henmon-Nelson and Otis group intelligence tests. The research sample had a mean IQ of 120 on the Stanford-Binet with a range of 98 to 153. The results indicated that of the 36 students predicted as gifted by the Otis, only 13 (36%) were identified as such by the Stanford-Binet. On the Henmon-Nelson 13 of 26 students (57%) were correctly predicted as gifted. Because 19% of the gifted students were not identified by either group test, both tests proved to be poor predictors of giftedness.

More recently, Harrington (1982) also found that group IQ tests tend to underestimate the IQs of many intellectually gifted students. According to Harrington, for every student identified as gifted on a group IQ test, one gifted child is not referred. Harrington suggested that the higher the ability level, the greater the discrepancy between individual and group IQ scores. He also found that a child's IQ may vary by as much as 30 points between group and individual tests. Further, because there may be a very small number of items at the greater difficulty levels on group tests, a child may have to perform perfectly to be predicted as gifted.

#### Intelligence Quotient Short Forms

So-called IQ short forms are comprised of abbreviated versions of individually administered standardized intelligence tests. Typically selected for short forms are subtests of the WISC-R or items from the Stanford-Binet. Short forms of the Wechsler Scales and Stanford-Binet have been studied extensively (Birch, 1955; Carleton & Stacey, 1954; Enburg, Rowley, & Stone, 1961; Findley & Thompson, 1958; Grossman & Galvin, 1987; Meister & Kurko, 1951; Nichols, 1962; Simpson & Bridges, 1959; Wright & Sandry, 1962; Wade, Phelps, & Falasco, 1986; Yakowitz & Armstrong, 1955; Zimet, Farley, & Dahlen, 1985). However, it is only since 1978 that short forms have been relatively widely

studied as a method for screening potentially gifted students.

An early attempt to predict gifted IQ using a short form test was conducted by Thompson and Findley in 1962. Finding that the Similarities (S), Information (I), Picture Arrangement (PA), Block Design (BD), and Picture Completion (PC) WISC subtests could be effectively used for this purpose, Thompson and Findley published the California Abbreviated WISC for the Intellectually Gifted (CAW-IQ) in 1966).

In their study, Killian and Hughes (1978) measured the effectiveness of the Lyman short form (Lorr & Meister, 1942) and the Vocabulary-Block Design subtests of the WISC-R dyad for predicting IQ on the Stanford-Binet and WISC-R respectively. Subjects were 142 students between 5- and 15-years-old possessing a mean IQ of 125. Results indicated a correlation of  $r=.92$  between the WISC-R and V-BD dyad whereas the Stanford-Binet and Lyman scores were correlated at  $r=.78$ . Killian and Hughes did not present results of the actual number of students correctly predicted as gifted. They did, however, indicate that 32% of the students had short form/Full Scale IQ score discrepancies of 6 points or more.

Employing a much larger sample of students than did previous researchers, Karnes and Brown (1981) used

Silverstein's (1970) method of deriving "the best short form combinations" (p. 169) to obtain an accurate gifted IQ predictor. Silverstein's method takes subtest unreliability into account when measuring predictive ability. Nine hundred, forty-six gifted children ages 6.0 to 16.0 ( $\bar{X}$  chronological age [CA] = 9.9) served as subjects. Karnes and Brown found that the WISC Block Design subtest was represented frequently in subtest combinations that correlated with WISC Full Scale IQ. Supporting Killian and Hughes' findings, the V-BD dyad was found to be the most accurate for predicting gifted IQ. The use of subtest tetrads was found to be useful, increasing correlation coefficients from .628 to .734. Again, actual accuracy ratios were not provided in the study.

Proceeding under the notion that, "since a short form IQ test is composed entirely of some subset of questions of items taken directly from a full-length IQ test, a short form would seem to be an ideal predictor of full-length IQ test performance" (p. 40), Dirks, Wessels, Quaforth, and Quenon (1980) compared various short form combinations with Full Scale IQ on the WISC-R. Subjects consisted of 47 fourth graders with a mean IQ of 123 (range = 106 to 144). Twelve WISC-R subtest combinations were studied. It was revealed that the short form combinations of Similarities, Object Assembly and Vocabulary and S-OA

were each good predictors of Full Scale IQ. Although correlations on the BD subtest were high, as shown in previous studies, they tended to predict an excessive number of nongifted students as gifted. The S-OA dyad predicted 8 of 11 gifted students and 4 who were not. The S-OA-V triad predicted 9 of 11 gifted students and 4 who were not.

Utilizing the studies by Killian and Hughes (1978) and Dirks et al. (1980), who noted that V-BD and S-OA dyads, respectively, were the most effective in predicting Full Scale IQ, Fell and Fell (1982) evaluated 92 WISC-R protocols of children previously evaluated as gifted program candidates. The students ranged in age from 6-0 to 11-7 ( $\bar{X}$  age = 8.4) and possessed Full Scale IQs of 130 or greater. Eleven subtest dyads were studied in terms of frequency with which each produced an estimated IQ  $\geq$  130. Greatest predictive accuracy was achieved using the S-V and S-OA dyads. These correctly predicted 62% as gifted. The I-BD dyad yielded prediction ratings of only 43%. Not providing an exact number, the authors indicated that some gifted children were overlooked. While results are consistent with findings by Dirks et al., indicating that the S-OA dyad is most effective, prediction accuracy was much lower in this study.

In a fairly recent study, Kramer, Markley, Shanks, and Ryabik (1983) utilized Thompson and Findley's (1966)

CAW-IQ on a sample of 73 children, ages 6-0 to 16-7 ( $\bar{X}$  age = 10-5). All subjects received the WISC-R and all were analyzed in terms of the S, I, PA, BD, and PC subtest pattern. Of the 48 students predicted as gifted, 39 were predicted accurately; of 25 students predicted not to be gifted, 21 were correctly described. This subtest short form was considered to be a relatively accurate predictor of gifted IQ.

#### Intelligence Quotient Screening Tests

Individually administered intelligence tests designed to estimate mental ability, usually in 20 minutes or less, have become a widely used procedure for screening gifted intelligence. The Slosson Intelligence Test (Slosson & Jensen, 1982) is one such screening test. It had been adopted in the Hillsborough County Florida school district for the purpose of screening the gifted. High correlations between the SIT and the WISC-R or Stanford-Binet have been reported in research findings (Lawrence & Anderson, 1979; Martin & Kidwell, 1977; Martin & Rudolph, 1972; Mize, Smith, & Callaway, 1979; Ritter, Duffy, & Fischman, 1973; Slosson & Jensen, 1982; Stewart & Jones, 1976). However, the few available studies conducted on gifted samples have not supported the use of the SIT for screening.



In the previously discussed study (see Chapter I) by Grossman and Johnson (1983), the Otis Lennon group IQ test was found to be a better predictor of gifted achievement than the SIT for high achieving students. Dirks, Wessels, Quaforth, and Quenon (1980) also found the SIT to be a poor predictor of gifted ability. They administered the SIT to 47 academically talented fourth graders. The students were also administered the WISC-R to determine their actual IQ scores. Of the 11 students who were found to possess gifted intelligence ( $IQ \geq 130$ ), only 8 were identified as such by the SIT. In addition, the SIT falsely predicted gifted intelligence in 9 of the 38 nongifted children. The researchers concluded that the SIT alone should not be used to predict IQs of gifted children. The SIT has also been found to significantly overestimate IQ scores. Machen (1972) investigated the reliability and concurrent validity of the SIT with the WISC, using 5 gifted children ages 9 through 11. The results revealed a significant correlation between the two tests, though the SIT tended to overestimate the WISC by at least one standard deviation. Additionally, the SIT has been shown to underestimate IQ scores. Mize et al. (1979) found, in their study of 207 students from all grade levels, that of students with above average intelligence, 24% were overestimated and 24% were underestimated by 11 or more IQ points on the SIT.



In 1979 Karnes and Brown further examined the tendency of the SIT to over- or underestimate IQ scores. In this study the validity of the SIT in relation to the WISC-R was assessed for a group of 79 gifted children ages 6 through 12. A SIT-WISC-R correlation of  $r=.48$  was calculated; this coefficient was significant at the .001 level. The authors also computed a regression equation with which to predict WISC-R IQ from the SIT. Results indicated that at the lower ranges of SIT scores, it tended to underestimate the WISC-R, while at the upper ranges IQ was overestimated. Despite the high correlation between the two tests, it is apparent that Karnes and Brown were not confident in the SIT's predictive ability because they recommended using two SEMs for the SIT when screening gifted IQ to ensure that most gifted students are identified. To obtain 95% accuracy, a cutoff score of 105 would have been necessary. However, Karnes and Brown did not indicate how many nongifted students would have been predicted as gifted using a cutoff this low.

In a similar study, presented at the Annual Meeting of the Alabama Association of School Psychologists in 1983, Apple discussed the precision of the SIT in predicting WISC-R IQ of 61 gifted students ages 6 to 11. Differences in scores were compared by use of independent t-tests. The results supported the findings of Karnes and

Brown that at the lower SIT ranges WISC-R IQs were underestimated and that at the upper SIT ranges the WISC-R IQs were overestimated. Apple concluded that valuable diagnostic information yielding a qualitative picture of the child's strength is omitted when the SIT alone is used as a screening indicator.

Whereas Karnes and Brown as well as Apple compared SIT and WISC-R scores for youngsters already placed in gifted classes, Rust and Lose (1980) attempted to accurately screen potentially gifted students in first through seventh grade. Based on teacher referrals and SIT scores of 130 or above, 438 students were found eligible for WISC-R evaluation. Of these, 132 were utilized in the research sample. According to stepwise regression equations, the SIT was found to be a significant predictor of Full Scale IQ. However, of the 132 students predicted, only 61 achieved WISC-R IQs of 130 or above. Thus, setting the SIT cutoff at 130 failed to screen out 54% of the nongifted students. If a cutoff of 134 had been used, as suggested by Karnes and Brown, 42 evaluations would have been eliminated. However, of those 42, 12 would have been gifted. Karnes and Brown noted that while there was a high correlation between the SIT and WISC-R, there was a great deal of variability with individual cases. It was concluded that in all studies high error can be expected

when the SIT is used to predict WISC-R IQ among the gifted.

Other IQ screening tests such as Guilford's Structure of the Intellect Test (SOI) (Pearce, 1983; Stenson, 1982), Ravens Progressive Matrixes (Pearce, 1983; Petty & Field, 1980), The Peabody Picture Vocabulary Test (Mize et al., 1979; Pedriana & Bracken, 1982), and The Ammons Quick Test (Joesting & Joesting, 1971; Kendall & Little, 1977; Nicholson, 1977) have been correlated with the WISC, WISC-R, and the Stanford-Binet. In some instances significant correlations have been found. However, very few studies have been conducted with samples of gifted students. In one such study, DeFilippis and Fulmer (1980) found that the Ammons Quick Test underestimated WISC-R IQ for 99 first, fourth, and seventh graders with high ability.

In another study involving samples of gifted students, Wright (1983) correlated WISC and Peabody Picture Vocabulary Test (PPVT) scores of 35 students referred by teachers for gifted program testing. A correlation of  $r=.27$  was calculated and it was found that nearly half of those who scored two standard deviations above the PPVT mean were not eligible for gifted program placement based on WISC-R IQ scores. Wright recommended that the PPVT not be used to screen gifted program candidates.

A third study was conducted by Stenson (1982) to determine the concurrent validity of the Structure of

Intellect (SOI) Gifted Screener with the WISC-R. The subjects were 3239 elementary school students. A multiple correlation of  $r=.337$  was significant at  $<.05$ ; however, only 11% of the variance in WISC-R scores was explained by the Gifted Screener. No predictor variable contributed to a significant multiple correlation coefficient when Full Scale IQ or any combination of WISC-R subtests was used as the criterion variable. Stenson concluded that the Gifted Screener should not be used to predict WISC-R IQ for gifted program prospects.

In 1985, Clarizio and Mehrens evaluated the technical data manuals for the SOI to determine the test's value as a screening test for gifted intelligence. It was concluded that "the SOI model has severe psychometric limitations" (p. 119). These limitations center around poor reliability, inadequate normative data, and poor external validity for many of the factors measured by the test.

### Achievement Tests

In research introduced in Chapter I it was suggested that achievement tests (CTBS) and cognitive ability tests (TCS, WISC-R, S-B) measure much the same construct. In the supportive literature were indications that distinctions between achievement tests and cognitive ability tests are often unclear. Correlational and factor analytic studies lend credence to this contention.

Lennon (1978) has found that relationships between intelligence and achievement tests are so strong as to lead to the criticism that the two types of tests do not measure anything different. Both tests measure what the student has learned (Gronlund, 1976) and both tests predict future learning with similar degrees of success. IQ tests and achievement tests differ in form but not necessarily in content (Mercer, 1979).

In a series of studies in the late 1970s and 1980s WISC-R IQs were correlated with achievement subtests of the Wide Range Achievement Test (WRAT). Consistently high correlations were found. Some of these early studies are summarized in Table 1-1.

Also, in 1978, Stedman, Lawlis, Cortner, and Achtenberg attempted to relate Kaufman's (1975) factor scores to WRAT attainment in a population of 76 children, ages 6 to 13. Correlations were found to be positive and significant.

Yule, Gold, and Busch (1981) administered the WISC-R and a battery of achievement tests to students at age 16 1/2. Achievement measures included tests of "sentence reading," spelling, and arithmetic. WISC-R, Verbal IQ shared 50% of the variance in reading, spelling, and arithmetic. Correlations between Full Scale IQ and achievement were as high as  $r=.80$ .

Table 2-1. Summary of Relationship between WISC-R and WRAT.

WISC-R	WRAT		
	Reading	Spelling	Arithmetic
Brooks (1977) N=30; 6-10 years			
V.S. IQ	64 <sup>a</sup>	55	74
P.S. IQ	71	70	71
F.S. IQ	70	65	76
Hartlage and Steele (1977) N=36; Mean age = 7 yrs 9 months			
V.S. IQ	75	35	76
P.S. IQ	54	33	67
F.S. IQ	68	35	76
Schwartz and Schwartz (1977) N=282; 6-16 years			
(a) 6-11 yrs			
V.S. IQ	68	61	69
P.S. IQ	63	60	69
F.S. IQ	72	65	75
(b) 12-16 yrs			
V.S. IQ	74	69	66
P.S. IQ	40	34	55
F.S. IQ	62	56	66
Hale (1978) N=155; 6-16 years			
V.S. IQ	54	49	64
P.S. IQ	29	26	44
Full Scale correlations not quoted.			

<sup>a</sup>Decimal points omitted.

In 1982 a follow-up to the studies summarized in Table 2-1 was conducted by Grossman and Johnson. In their study, 77 students ages 6 to 16 were administered the WISC-R and the WRAT. Factor scores were computed on two of Kaufman's (1975) factors (Verbal Comprehension and

Freedom from Distractibility) and WRAT subtests. A multiple regression analysis was computed wherein WISC-R factor scores served as conjoint predictors and the WRAT standard scores were employed as criterion variables. Results indicated a significant overall prediction of WRAT reading, spelling and arithmetic by the two WISC-R factors.

Wright and Dappan (1982) assessed 250 students with a mean age of nine years on the WISC-R and WRAT. Factor analysis showed a common factor for subtests on both measures. Correlations between subtests from the two tests were as high as  $r=.60$  (on WISC-R, Arithmetic and WRAT, Arithmetic). Some other subtests correlated at .40 to .50.

Literature concerning the overlap of individual tests of intelligence and tests of achievement include studies in which the WISC-R and the Peabody Individual Achievement Test (PIAT) (Dunn & Markwardt, 1970) were examined. Wikoff (1978) factor analyzed the WISC-R along with the PIAT for 180 referred children. Although the PIAT General Information and Mathematics subtests loaded on factors previously identified in the structure of the WISC-R, the remaining subtests loaded on a separate factor, subsequently labeled Word Recognition. The results supported the use of both instruments as sources of mutual but



supplementary information in the assessment of learning problems.

Dean (1977) assessed the degree of redundancy between the WISC-R and the PIAT using a canonical correlation analysis with scores from 205 referred children. The results indicated that 65% of the functions of the PIAT overlapped with the WISC-R and that 37% of the functions of the WISC-R overlapped with the PIAT. The overlap was attributed to common verbal-educational content. Dean (1982) found a similar asymmetrical overlap between these measures in samples of 100 Anglo and 100 Mexican-American children. As in Wikoff's factor analysis, both of Dean's analyses showed the PIAT subtests of reading and spelling to offer the greatest degree of information not redundant with the WISC-R.

Brock (1982), finding that a paucity of research existed for factor analytic investigations of the WISC-R in combinations with individual achievement tests, conducted such a study. He factor analyzed the WISC-R, WRAT, and PIAT for 183 male students in grades 3 through 6. An attempt was made to determine the traits or common skills measured by IQ and achievement tests when viewed concomitantly. Four factors emerged. One, a numerical factor, was comprised of subtests from all three tests.



Moderately high correlations  $r=.40$  to  $.50$  were found between some of the other IQ and achievement subtests.

Stewart and Morris (1977) factor analyzed the WISC, WAIS, WRAT, and CAT (California Achievement Test) for 182 students ranging in age from 11 to 18. A "substantial" overlap of verbal intelligence and academic achievement was found. Resulting factors conformed reasonably well to those of Kaufman (1975). Subtests from each measure were found to load on each of the IQ factors.

In a study wherein the abilities underlying reading readiness were identified, Olsen and Rosen (1971) factor analyzed three group reading tests and the WISC. Subjects consisted of 218 first graders. The 35 subtests were correlated and the resulting matrices subjected to a principal component analysis. Four common factors were revealed. In one factor, reading comprehension loaded with four WISC-R subtests. In another, "writing letters" correlated highly with WISC-R Vocabulary. In a third factor, sentence writing and WISC-R Coding were included.

There have emerged two camps of thought on the issue of reading skill acquisition and intelligence. On one hand, in some research it has been suggested that reading is a function of information processing or encoding skills as opposed to being a primarily intellectual function. On the other hand, in similarly focused factor analytic investigations reading has been found to be highly loaded

on a general intelligence factor and a good predictor of intelligence.

Researchers whose views represent the latter viewpoint have supported the notion that reading ability is highly correlated with general intelligence and is a good predictor of intelligence. Jensen (1981) reported a correlation of  $r = .68$  between reading comprehension and Full Scale IQ for a large sample of students. Other researchers have cited similarly high (.60-.70) correlations between reading and IQ for various samples of students in grades K through 12 (Brooks, 1977; Hale, 1978; Hartlage & Steele, 1977; Ryan, 1979; Wikoff, 1978; Yapp, 1977; Yule, Gold, & Busch, 1981). In a literature review of 34 studies, Hammill and McNutt (1981) found a median correlation of .75 between measures of intelligence and achievement.

Reasoning that the most efficient gifted screening assessment would be significantly correlated with achievement if giftedness is defined as superior school-related ability, Grossman and Johnson (1983) investigated the Stanford Achievement Test. They found a significant correlation with intelligence among 46 children with SIT IQs above 120.

In another pertinent study, Schena (1963) found that of 226 sixth and seventh graders who scored two or more

"levels" above the norm on the Metropolitan Reading Test, 61% scored above 130 on the Stanford-Binet. In his 1984 study, Sternberg found that IQ accounted for as much as 25% of the variance in scholastic performance.

In two of the few other studies in which achievement was correlated with intelligence among superior students, Mayfield (1979) had 573 third graders evaluated in terms of intelligence, achievement, creativity, and teacher perception of IQ. Results yielded significant correlations between intelligence and a wide range of achievement domains among the student sample. Similarly, Karnes, Edwards, and McCallum (1986) found a significant correlation between total scores on the California Achievement Test (CAT) and WISC-R Full Scale IQs of 41 gifted children in grades four through six.

Thus the results of this body of literature comparing intelligence with achievement appear conclusive, as in a substantial number of studies it is suggested that the two variables are fairly highly correlated.

Mallinson (1963) attempted to uncover a relationship between intelligence and achievement in science and math. The SRA achievement series and the SRA Primary Abilities Test were given to secondary grade students. There was a resulting correlation of  $r=.65$  between verbal ability and science (facts and principles). Verbal ability was also

found to have reasonably high correlations with factors of arithmetic achievement.

In another study in which factors related to math performance were investigated, Roach (1979) reported a significant correlation ( $r=.80$ ) between arithmetic achievement and verbal IQ in third graders.

Correlation coefficients for CTBS and TCS subtests (CTB/McGraw-Hill, 1984) were calculated using 2813 seventh graders. Coefficients between .60 and .72 were not uncommon. Correlations between the TCS Total Score and CTBS subscales of Reading, Language, Math, Social Studies and Science were .71, .71, .68, .65, .71 and .65, respectively. The CTBS and TCS Total Batteries correlated at  $r=.75$ . These high correlations suggest that both tests may measure similar, though operationally distinguishable, constructs. Support for the contention that these tests represent similar constructs may also be found in the correlations of TCS subscale scores with those scores of its predecessor, the Short Form Test of Academic Aptitude (SFTAA). The range of correlations was .55 to .82, not dissimilar to those of the TCS and CTBS. The fact that the average correlations were positive means that the subscales must be measuring something in common (Jensen, 1980).

In summary, there is substantial evidence that achievement test scores and IQ test scores are highly

correlated. The research also provides reason to believe that a common factor underlies performance on both types of tests.

### Summary

In the literature relevant to the accuracy of various procedures for screening intelligence among gifted students, there have generally been mixed results. Some encouraging findings have occurred on studies of achievement ratings and short-form IQ tests. Group IQ tests have tended to underestimate the IQs of some gifted children, though they generally predicted gifted IQ with moderate accuracy. Much less effective means for predicting gifted IQ are IQ screening tests.

The preceding literature has focused on the problems of screening gifted intelligence among the school age populations. Those procedures that have shown some success have been inconsistent in their findings. Nearly all have proved inefficient in terms of time and cost.

### Instruments Used in Study

#### Wechsler Intelligence Scale for Children-Revised (WISC-R)

The WISC-R has been the most widely administered test of children's intelligence (Bryan & Bryan, 1975; Grossman

& Galvin, 1987; Salvia & Ysseldyke, 1978; Vandiver & Vandiver, 1979). In much of the research surrounding this instrument it is suggested that it merits this distinction. Friedes (1978) described the standardization of the WISC-R as "state of the art" (p. 232) and as meriting "blue ribbons." In addition, he noted as praiseworthy the high correlation coefficients between the WISC-R and the Stanford-Binet.

Reliability coefficients of internal consistency for the WISC-R Verbal, Performance, and Full Scale IQ scores reported in the test manual were obtained by utilizing a formula for computing reliability of a composite group of tests (Wechsler, 1974). The average reliability coefficients across the range of age levels were V,  $r=.94$ ; P,  $r=.90$ ; and FS,  $r=.96$ . The coefficients of individual subtests based on split-half or test-retest methods ranged from  $r=.77$  to  $.86$ . Test-retest correlations for the Verbal, Performance, and Full Scale IQs ranged from  $r=.90$  to  $.95$  based on a 3-month interval between tests.

Factor analytic research by Kaufman (1979) has shown that factors corresponding closely with the Verbal and Performance Scales of the WISC-R exist. In 1980, Karnes and Brown factor analyzed the WISC-R on 946 gifted students ages 6.0 to 16.0. The resulting factors were consistent with those found by Kaufman on the normal population. Most verbal scale subtests had factor loadings in

Perceptual Organization. These studies strongly support the validity of the WISC-R.

### Slosson Intelligence Test

The SIT is an IQ test for children and adults designed for use by either relatively untrained examiners or qualified professionals. The SIT typically takes between 10 and 30 minutes to administer.

New norms (1982) represent a significant departure from procedures previously employed (1961) for calculating an IQ score. In norming the SIT, the Stanford-Binet was used as the anchor test. Consistent with the 1974 revision of the Stanford-Binet, ratio IQs were abandoned in favor of deviation IQs. Frequency distributions were calculated for each of the 19 chronological age ranges on both IQ scales. Then, utilizing a "modified table look-up approach," appropriate IQs from the Stanford-Binet were entered on the developing SIT tables. The mean IQ for the SIT is 100 and the standard deviation is 16.

In the SIT manual, the authors present evidence to persuade the reader that the revised SIT IQs are equivalent to Stanford-Binet IQs. This task is undertaken, in large part, by comparing previous (1961), less positively correlated coefficients to newer data.

Slosson and Jensen (1982) stated that "the SIT is as accurate as the Stanford-Binet in measuring a person's



intelligence when both instruments have been properly administered" (p. 16) and further proposed that the SIT qualifies as an alternate form "of the Stanford-Binet because the two tests possess equivalent means and standard deviations" (p. 16). Based on their dubious assumption of test equivalency between the revised SIT and the S-B, the authors employed the Mean Absolute IQ Difference (MAD) statistic to determine alternate form test reliability and standard errors of measurement. The MAD procedure, which is meant to be used only with equivalent forms of a test, yields a statistic which is approximately equal to the standard deviation times  $\frac{1}{\sqrt{2}}$ . The authors did not indicate the relative effects on the reliability and the SEMs when the measures compared do not strictly meet the criteria of alternate forms, as is apparent in this case. Nor are there attempts to evaluate other kinds of test reliability. It might be concluded, therefore, that the authors' claim that "the SIT's reliability may be regarded as not less than .95" (p. 136) should be interpreted with caution.

Another claim made in the SIT manual is that the mean difference between the Stanford-Binet and Slosson IQ scores is less than one point, based on the sample of 1,109. The procedure by which this statistic was obtained entails computing the means for IQ differences between the

two tests for three IQ levels across four age groups: below 84, 84-116, and above 116. For example, at age 13-6 and above, mean differences between the Stanford-Binet and SIT are -1.41, -1.10, and 2.62 at the three IQ levels. In calculating the mean difference, negative means are added to positive means resulting in a misleadingly low overall mean difference. In this example the total mean difference for age 13-7 and above is -.67. However, if individual means had been summed in terms of nondirectional deviation from zero, the mean difference would have been approximately 1.7. With regard to the mean difference for the entire sample, when the nondirectional procedure is used the difference changes from -.04 to approximately 1.4. The mean scores are rendered even more difficult to interpret because no standard errors for the means are reported.

In spite of the apparent inconsistencies in the new SIT manual, the revisions, particularly in its renorming, represent considerable improvement in the test's validity and reliability. These test improvements, along with the ease of its administering and scoring, render the SIT a test of considerable utility as an intellectual screening procedure.

Comprehensive Test of Basic Skills (CTBS)/Test of  
Cognitive Skills (TCS)

The CTBS (CTB/McGraw-Hill, 1984) and TCS (CTB/McGraw-Hill, 1984) are the tests from which NP items were taken. Psychometrically, these tests were well suited to this research. The appropriateness of the CTBS and TCS for this research are supported by several of their attributes, some of which were discussed in Chapter I.

Items were chosen for both tests according to item response theory (IRT) utilizing a three-parameter logistic model. The items were chosen according to their ability to (a) discriminate high ability traits from low ability traits, (b) discriminate high ability students and low ability students by matching item difficulty with student total score, and (c) account for guessing as an influence on score difficulty.

In terms of content validity, the CTBS is designed to measure understanding of a broad range of concepts as developed by various educational curricula. Test performance reflects a student's skills in effective use of information explicit in categories derived from Bloom's taxonomy (Bloom, 1956). Item development specifications were designed to ensure comprehensive coverage of the content and process categories.

The TCS is designed to measure an aptitude construct that can be operationally distinguished from the

achievement construct of the CTBS, based on research conducted at McGraw-Hill by Buchet (1974, cited in TCB/McGraw-Hill, 1984). Empirical criteria for distinguishing between aptitude and achievement measures were derived by the publishers.

Product moment correlations between the four subtests of the TCS were between  $r=.41$  to  $r=.65$ . Coefficients between subtests and total score ranged from  $r=.72$  to  $r=.85$ . Therefore, it was suggested that all subtests measure general intelligence but also measure independent factors. A correlation coefficient between the CTBS and TCS of  $r=.78$  was calculated on a sample of seventh graders.

Another attribute of the CTBS and TCS is the comprehensive sampling and norming standards applied. The norming samples contained approximately 250,000 students in grades K-12 from public, Catholic, and other private schools (CTB/McGraw-Hill, 1984). School districts were randomly chosen from four geographic regions. Comprehensive norming and standardization information is available in the CTBS and TCS Technical Reports.

Internal reliability coefficients were calculated according to the Kuder-Richardson formula 20. CTBS reliability coefficients ranged from .30 to .96 on the 10 subtests (CTB/McGraw-Hill, 1984). All subtests except Spelling and Reference Skills had values at or above .90. On the four TCS subtests reliability coefficients ranged from

.80 to .84. The TCS Technical Reports provide reliabilities on SEMs for subtests based on number correct at each grade level. Composite calculations for the total test are not provided. Also reported are bias studies and tables indicating how test biases are accounted for and controlled.

In summary, the CTBS and TCS were well suited for this study because of the sophisticated method utilized in analyzing items and the tests' high validity and reliability. Further, both tests employed sampling procedures designed to provide norms for the entire U.S. school population. Research has also been conducted to aid in reducing test bias for the CTBS and TCS.

#### Stanford-Binet Intelligence Scale (S-B)

The third revision of the Stanford-Binet (S-B), published in 1960, remained unchanged in content and format through 1985. A revised version of the S-B was published in 1986. The 1960 version was constructed by combining forms L and M of the 1937 scale and eliminating those items considered obsolescent and by relocating items whose difficulty level had altered during the intervening years. The test was, however, restandardized in 1972. New norms were derived from a sample of approximately 2,100 cases during the 1971-72 school year. Children in the 1972 norm group were chosen from 20,000 school age children in

grades 3 through 12 who were identified based on scores from the Cognitive Abilities Test. The distribution of scores in this subsample corresponded to the national distribution of the entire sample. The 1972 norms were believed to be based on a more representative sample than previous norms (Terman & Merrill, 1973).

The reliability of the 1937 Stanford-Binet was determined by correlating IQs on forms L and M administered to the standardization group within an interval of one week or less. Such reliability coefficients are thus measures of both short term temporal stability and equivalence across the two item samples. In general, the Stanford-Binet tends to be more reliable for older than for younger age groups, and for lower than for higher IQs (Anastasi, 1976). Reliability coefficients range from .83 to .98. The Stanford-Binet is considered a highly reliable test with most coefficients for the various age and IQ levels being over .90.

Validity ratings for the Stanford-Binet were obtained from examination of the test content, from factor analysis, and from correlations with achievement ratings. An examination of the Stanford-Binet tasks indicates assessment of a wide range of reasoning abilities. These include tasks requiring hand-eye coordination, perceptual discrimination, arithmetic reasoning, and verbal

reasoning. The most common type of test, especially of the upper age levels, is that employing verbal content.

Data on criterion-related validity of the Stanford-Binet have been obtained chiefly in terms of academic achievement (Anastasi, 1976). Correlations between the scale and school grades, teachers' ratings, and achievement test scores generally fall between .45 and .75. The Stanford-Binet tends to correlate highly with performance in nearly all academic courses, but predominantly with verbal courses such as English and history. Correlations with achievement test scores show the same pattern. The rigorous standardization and renorming of the Stanford Binet, along with its high validity and reliability, indicate that it was an appropriate IQ test for this study.



## CHAPTER III METHODOLOGY

### Overview

In this research study a procedure was investigated for analyzing a comprehensive, group-administered achievement and cognitive abilities test to determine whether an item set can be derived that discriminates gifted from high-achieving, but not-gifted, seventh graders. When such an item set was derived, it was compared to a commonly used IQ screening test to assess the relative accuracies of each procedure in discriminating gifted from not-gifted students in a second seventh grade population sample.

This chapter is organized into the following sections: (a) Population and Sample, (b) Assessment Instruments, (c) Research Procedures, (d) Data Analysis, and (e) Methodological Limitations.

### Population and Sample

The research sample of 179 students was drawn from a population of seventh graders who had been tested on the WISC-R or S-B for the "gifted program" in the Hillsborough

County (Florida) Public School District. Sampling was conducted at the end of the 1984-85 school year. Nearly all students had previously been administered the SIT. In most instances, only those students who scored two standard deviations or more above the mean had been administered the WISC-R or Stanford-Binet. All students in the population also had current CTBS/TCS scores on file. Some students in this population had met intellectual eligibility guidelines for the gifted program (on WISC-R or S-B criteria) and some had not. All students were tested by school psychologists during each of the three school years under investigation.

Simple random sampling was conducted by the researcher at the Hillsborough County School Board office in June of 1986. Names of the seventh graders who were tested for the gifted education program from September of 1982 and June of 1985 were obtained from computer printouts containing data for all students in the district who had been tested by school psychologists. Students in the sample pool were assigned a number, and numbers were selected according to a random number table. Numbers were then recorded and returned to the pool to ensure an equal chance of selection for the remaining numbers. After 61 students were randomly selected for Phase II, the remaining 118 students were assigned to Phase I.

### Assessment Procedures

As previously discussed, the five assessment instruments used for this research were the Wechsler Intelligence Scale for Children-Revised (WISC-R), The Stanford-Binet Intelligence Scale (S-B), the Slosson Intelligence Test (SIT), the Comprehensive Test of Basic Skills (CTBS), and Test of Cognitive Skills (TCS).

Administration, scoring, and interpretation of the WISC-R and Stanford-Binet were conducted by state certified school psychologists prior to the onset of this study. All tests were individually administered and hand scored using current norms. The WISC-R yields a Verbal Scale IQ (representing verbal reasoning abilities), a Performance Scale IQ (representing perceptual organization and nonverbal reasoning), and a Full Scale IQ. Only the Full Scale score, which represents total IQ, was used as a measure of gifted intelligence. This procedure conformed to school district guidelines. The Stanford-Binet yields a total IQ score only. A cutoff score of 127 was used as the gifted cutoff in the district. Students attaining a Full Scale IQ of 127 or greater were considered to have met the intellectual criterion for gifted program eligibility. The IQ cutoff was chosen by the district as a score that is two standard deviations above the test mean (WISC-R IQ=130), minus one standard error of measurement (three IQ points) (S-B IQ = 132 minus 5 IQ points). The

Slosson Intelligence Test was administered to students as an IQ screening procedure by school guidance counselors or curriculum specialists who typically had little formal training in administration of individual intelligence tests. SITs were given to students within one year prior to WISC-R testing. SIT protocols were hand scored by the test administrators. Under usual circumstances, a total IQ of 135 was used as a screening cutoff. Children who scored at or above this cutoff point were normally referred to the school psychologist for WISC-R evaluation. There were some exceptions to this rule because occasionally students who did not score at or above the cutoff were referred. Generally, these students exhibited extremely high academic skills or other competencies that compelled school personnel to refer them for formal testing. The SIT yields a total IQ score. The 135 cutoff score is two standard deviations above the test mean.

The CTBS and TCS were administered to students by classroom teachers in group format according to standardization procedures found in the teacher's manual. The CTBS and TCS were designed to be easily administered (Ahmann, 1972), and teachers have received little formal training in their administration. The seventh-grade level of the CTBS (Level H) yields subscale scoring in Reading (two sections), Spelling, Language (two sections),

---

Mathematics (two sections), Reference Skills, Science, and Social Studies. The Science and Social Studies subscales, consisting of 86 items, were not administered in all schools of the population school district; however, these data were included in this research. In addition to subscale scores, the CTBS yields an overall achievement index.

The TCS is comprised of four cognitive ability subtests measuring competencies in Sequencing, Analogies, Memory, and Verbal Reasoning. Derived scores are provided for subtests and for the overall profile. Because in this study responses to individual test items were analyzed, derived subscale and total scores for the CTBS scores were not utilized.

### Research Procedures

In Phase I of this study items comprising the new procedure (NP) were selected from the CTBS and TCS by conducting item analyses of the performance of the 118 students in the Phase I sample. Two sets of items were delineated that, in general terms, were answered correctly by the gifted students more frequently than by the not-gifted students.

Gifted cutoff scores on the obtained NP items were computed by subtracting fractions or multiples of standard deviations from the mean NP score of the gifted group to

determine which cutoff point(s) most accurately predicted giftedness. Newly referred students in the school setting who scored at or above that point would be referred for formal IQ testing. By adopting a cutoff score two standard deviations below the gifted mean, approximately 97% of intellectually gifted students would be referred for IQ testing based on research sample parameters. A cutoff score incorporating a one standard deviation below the mean cutoff would delete approximately 16% of the gifted students from testing and gifted program eligibility. However, a pitfall of including as broad a range of students as permitted by the two standard deviations criterion was that a relatively large number of not-gifted students would also be referred for formal IQ testing, thus reducing the accuracy of NP predictions of giftedness. Adjustment of the cutoff point was desirable to ensure an optimal ratio of gifted students accurately predicted to not-gifted students inaccurately predicted.

After items for the NP were obtained and multiple cutoff points was established, Phase II was begun. In Phase II, test scores based on selected NP items were computed for the second sample of students. The NP was analyzed in terms of its discrimination between the gifted and not-gifted students on WISC-R. The accuracy of classification was then compared to accuracy of classification

obtained using students' SIT scores as a screening procedure in predicting gifted IQ on the WISC-R.

Data collectors and recorders consisted of the researcher and employees of the district school board. School board employees working in the testing and evaluation office obtained CTBS/TCS item responses from computer data. WISC-R and SIT scores were obtained in a similar manner by the researcher and research assistant.

### Data Analysis

As discussed in Chapter I, two correlational analyses were conducted in Phase I of this study to determine desirable items for the NP. Phi coefficients, which are designed to correlate two dichotomous variables, were computed. The two dichotomous variables correlated were the student's item response (correct or incorrect) and student classification (gifted or not-gifted) on the WISC-R. The criterion for deciding if an item was to be included was significance at the .05 level. For the present sample, this meant that any item with a phi coefficient greater than .182, was selected for the NP screening test.

Phi is based on the proportions of cases passing and failing an item in both the gifted and not-gifted criterion groups. The phi coefficient is known to be biased



toward middle difficulty levels of test items. As previously discussed, the research (CTBS and TCS) items were designed primarily to assess performance of medium difficulty.

Test items were also analyzed according to the index of discrimination (Ebel, 1965). The difference between the percentage of gifted students and not-gifted students passing each item provides an index of item validity that can be interpreted independently of the size of the particular sample in which it was obtained (Anastasi, 1976). The index of discrimination (D) has been shown to measure item validity with equivalent accuracy to other more elaborate measures (Engelhart, 1965). Similar to the phi, D values are biased in favor of items with intermediate difficulty levels. A coefficient of .20 or greater was used as the criterion for selecting an item for the NP. Thus in Phase I, two forms of the NP screening test were created; one form, here designated as NP-phi, was based on items selected using phi coefficients; the other form, NP-D, was based on items selected using the index of discrimination.

For Phase II, analyses using the coefficient Kappa were conducted to test (a) whether the NP or SIT was more accurate in classifying gifted seventh graders, and (b) at what cutoff points either the NP or SIT was more accurate. The Kappa analysis measured the proportion of correct (and

incorrect) classifications for NP-phi, NP-D, and the SIT while adjusting for the percentage of correct classifications that could be expected on the basis of chance alone. Kappa is a descriptive statistic and not a test of statistical significance.

Using the Kappa statistic proportions of predictions were compared at various cutoff points. Kappa adjusts for predictions expected by chance alone by taking into account both observed and expected proportion classifications (Cohen, 1960). For example, when considering K values for the NP, the numerator of K is regarded as the proportion of students consistently classified by both the NP and the WISC-R (observed) over and above the product of the proportions of students classified individually by each test (expected). The denominator of K is the maximum possible increase in decision-consistency above chance level, given the proportions classified by the two tests independently.

The formula used to compute Kappa was

$$K = \frac{P - P_C}{1 - P_C}$$

where P = proportion of consistent gifted and not-gifted classifications for WISC-R and NP or WISC-R and SIT

and  $P_C =$  proportion of gifted classifications for WISC-R  
 x proportion of gifted classifications for NP  
 (or SIT) + proportion of not-gifted classifica-  
 tions for WISC-R x proportion of not-gifted  
 classifications for NP or SIT.

### Methodological Limitations

Possibly the most severe methodological limitation of this study concerns the appropriateness of the CTBS and TCS for discriminating students who score in a restricted range near the ceiling of the test. The CTBS/TCS tests were deemed as appropriate for this research because the data were readily available and, if usable, would preclude students from taking a gifted screening test. Also, the CTBS and TCS measure the wide range of skills and abilities. However, because the CTBS and TCS were designed to measure traits of the general population of students, they were less sensitive to group differences in the extreme ranges of ability and particularly at the ceiling level. The location parameter indicated the ability level, in scale score units, at which the item was most sensitive to individual differences. Thus, test items were designed to have their greatest sensitivity to individual differences in the general range where most students taking the test

would score. Since on most of the items, students in the research sample were expected to score well above the location parameter, most of the items provided little differentiation among these students with much higher scale scores. However, in defense of these instruments for use in this study, some items with extreme location parameters were purposely included in the CTBS by its constructors. These items were well above or below the range of performance for which the test was designed. Item characteristic curves indicate the existence of CTBS items that were passed typically only by students with very high total scores on the test (CTB/McGraw-Hill, 1984). IRT item location parameters for the TCS had not been calculated, though since the TCS was designed to predict achievement for the general population, it might be assumed that location parameter criteria would adhere to a similar rationale.

A second limitation of this study concerns measurement error due to the variable adherence to test standardization by administrators of the CTBS, TCS and the SIT. Because teachers and counselors receive varying degrees of formal training on the importance of precise adherence to test standards and on the influence of standardization deviations on test reliability, administrators' strict conformity to test standards was questionable. Because very little training was provided for the CTBS/TCS and SIT,

misinterpretation of instructions or standards was possible as well.

A related source of potential measurement error existed due to the group format by which the CTBS/TCS is administered. As opposed to the WISC-R and SIT, administrators were restricted in their ability to closely monitor individual students and control for such factors as misinterpreted directions or acute physical or emotional liabilities of students. Since regular CTBS/TCS testing was conducted only once a year in the research school district, efforts were made to test as many students as possible during that time. However, SIT and WISC-R testing were more easily postponed to a later date if a situation warranted such action.

Another limitation of the CTBS specific to this study was that some students were not administered the Science and Social Studies subtests in the school district, because they were optional and were administered only in some schools. These subtests are comprised of 40 items each and may contain items useful for discrimination of gifted and not-gifted students. If deleted from analysis, these items may detract from the overall accuracy of the proposed new procedure in accomplishing its intended goal.

The administration of the WISC-R and SIT over a 3-year period suggests a question regarding score

equivalence. One might argue that since the items administered to a 9-year-old and an 11-year-old are different and represent different test difficulty levels, the two students are, in fact, being tested on different scales. This argument implies that gifted intelligence for the 9-year-old is not equivalent to gifted intelligence for the 11-year-old. To the contrary, while imperfections in test stability will cause some fluctuation in IQ over time, the adoption of the deviation IQ (Wechsler, 1974) permits comparison of scores over age levels.

## CHAPTER IV RESULTS

This chapter is presented in a format that sequentially reflects the methodological progression of the study. Phase I will be discussed first to present results of two analyses used to select items for the new screening procedure (NP). Next, Phase II cross validation results are presented in terms of prediction accuracy of the NP-phi, NP-D, and SIT in relation to cutoff scores.

### Phase I--Item Selection

#### Phi Coefficients

The phi coefficient analysis of the 460 CTBS and TCS items yielded 56 items that discriminated the 59 gifted and 59 not-gifted seventh graders. This number of items is over twice as many as would be expected by chance at the  $\leq .05$  level of confidence. There were 13 items significant at the  $\leq .01$  level and three significant at  $\leq .001$  (Table 4-1). As also shown in Table 4-1, all CTBS and TCS subtests contributed items except for CTBS Reference Skills and TCS Sequences.

Not all of the 56 significant phi analyses items were retained for use in Phase II. In five cases the phi items



Table 4-1. Item Phi Values and Significance Levels

Item No.	Subtest	Phi-Value	Significance
16	CTBS-Voc.	.2059	$\leq .05$
39	Voc.	.3258	$\leq .001$
40	Voc.	.1834	$\leq .05$
41	Voc.	.2069	$\leq .05$
49	Read.	.1963	$\leq .05$
53	Read.	.2172	$\leq .05$
61	Read.	.2266	$\leq .05$
65	Read.	.1977	$\leq .05$
77	Read.	.1963	$\leq .05$
80	Read.	.2502	$\leq .01$
83	Read.	.2386	$\leq .01$
85	Read.	.2502	$\leq .01$
109	Spell.	.2645	$\leq .01$
113	Spell.	.2285	$\leq .01$
120	Spell.	.2199	$\leq .01$
122	Lang.	.2652	$\leq .01$
127	Lang.	.1842	$\leq .05$
129	Lang.	.2069	$\leq .05$
136	Lang.	.2187	$\leq .05$
137	Lang.	.2377	$\leq .01$
139	Lang.	.2559	$\leq .01$
154	Lang.	.2035	$\leq .05$
159	Lang.	.1842	$\leq .05$
171	Lang.	.1913	$\leq .05$
176	Lang.	.2934	$\leq .01$
177	Lang.	.2035	$\leq .05$
180	Lang.	.2018	$\leq .05$
182	Lang.	.1842	$\leq .05$
187	Lang.	.1913	$\leq .05$
193	Lang.	.1834	$\leq .05$
205	Math	.2187	$\leq .05$
215	Math	.3007	$\leq .01$
*216	Math	.1905	$\leq .05$
220	Math	.3245	$\leq .001$
230	Math	.3564	$\leq .001$
236	Math	.2652	$\leq .01$
264	Math	.2161	$\leq .05$
266	Math	.2331	$\leq .05$
270	Math	.2146	$\leq .05$
271	Math	.2806	$\leq .01$
273	Math	.2784	$\leq .01$
276	Math	.1885	$\leq .05$
277	Math	.2188	$\leq .05$
(2)311	Sci.	.3214	$\leq .05$

Table 4-1. Continued.

Item No.	Subtest	Phi-Value	Significance
* (2) 343	Soc. St.	.2812	$\leq .05$
(2) 364	Soc. St.	.3076	$\leq .05$
(2) 366	Soc. St.	.3921	$\leq .01$
(2) 378	Soc. St.	.3076	$\leq .05$
407	TCS-Anal.	.1977	$\leq .05$
* 421	Mem.	.2148	$\leq .05$
425	Mem.	.1916	$\leq .05$
429	Mem.	.2035	$\leq .05$
434	Mem.	.1835	$\leq .05$
* 438	Mem.	.2068	$\leq .05$
* 440	Verb. Reas.	.1858	$\leq .05$
457	Verb. Reas.	.2168	$\leq .05$

\* Item discriminates in favor of not-gifted.

2 Answered by less than 90% of sample.

discriminated in favor of the not-gifted students. That is, not-gifted students responded correctly to the items more frequently than the gifted students. Also deleted from Phase II analyses were five items responded to by only about half of the Phase I sample. These five items were located in the Science and Social Studies subtests, which were administered on an optional basis at the discretion of the various schools involved. They were deleted because the cutoff scores were set based on total items administered to all students. Those items not administered to all students were deleted so that all students could potentially attain the maximum raw score. Therefore, 47 phi items were retained for use in Phase II, 42 from the CTBS and 5 from the TCS.

The mean of the total scores for the 47 items selected by the phi analyses was 39.37. The median and mode were somewhat higher, 41.00 and 46.00, respectively. The standard deviation was 7.06 and the range was 41. Most students performed well on these items, with most responding correctly to nearly all of them. This homogeneity of scores is reflected by the negative skew of the distribution. Even though scores appeared to concentrate near the upper end of the distribution, the large range of scores (i.e., 4 through 46) contributed to a standard deviation of adequate size.

### Index of Discrimination

The index of discrimination analyses, which measured the difference between the percentage of gifted and not-gifted students passing each item, yielded a subset of 24 items with D values of .20 or greater. As recommended by Engelhart (1965), items possessing a D value of at least .20 are considered to show adequate discrimination. All 24 of the acceptable items were obtained from the CTBS.

In Table 4-2 the upper and lower values represent percentage passing each item for the gifted and not-gifted students, respectively. D values ranged from .221 through .393. As might be expected, the highest percentage of correct responding occurred with the beginning TCS subtest items. This was because TCS items are ordered hierarchically by subtest according to difficulty level.

In this analysis, also, an item was deleted if it discriminated in favor of the not-gifted group or if it was not administered to over 10% of the sample. In contrast to the phi analyses, a large proportion of items (i.e., 10 or 42%) were deleted. Fourteen items (Table 4-2) were retained from the CTBS.

In contrast with the phi data, results of scores based on selected items from the D analyses revealed a relatively normal distribution, with the mean of 10.02, median of 10.02, and mode of 11.00 falling within a range of one test item. The standard deviation was 2.62 and the

Table 4-2. Items with D Values  $\geq .20$  and Corresponding Subtests

Item No.	Subtest	Upper $\bar{x}$	Lower $\bar{x}$	D
39	CTBS. Voc.	.930	.667	.263
80	Read.	.912	.712	.200
85	Read.	.912	.712	.200
109	Spell	.931	.731	.200
158	Lang.	.958	.867	.208
176	Lang.	.897	.650	.327
185	Lang.	.879	.667	.212
215	Math	.842	.567	.275
230	Math	.873	.552	.323
271	Math	.948	.746	.202
273	Math	.947	.746	.201
277	Math	.737	.525	.212
(2)281	Math	.750	.538	.212
(2)311	Sci.	1.000	.735	.265
317	Sci.	.937	.727	.210
(2)330	Sci.	.937	.727	.210
(2)341	Soc. St.	.875	.667	.208
(2)343	Soc. St.	.688	.909	-.221*
351	Soc. St.	.937	.727	.210
(2)360	Soc. St.	.937	.467	.240
(2)364	Soc. St.	1.000	.435	.242
(2)366	Soc. St.	.938	.506	.393
(2)377	Soc. St.	.937	.727	.210
(2)378	Soc. St.	1.000	.758	.242

\*Item discriminates in favor of not-gifted.

2 answered correct by less than 90% of the sample.

range was 14. Unlike the phi distribution, the ceiling of the D distribution was probably sufficiently high because correct responding diminished beyond raw scores of 11. Only 11 of 117 students received a raw score above 12, while 65 received scores within one point of the mean. Taking into account the small number of NP-D items (14), the standard deviation of 2.6 is considered to be adequately large for calculating cutoffs. The range of scores covered both extremes of the distribution, and the skew is not as great as might be expected given the generally restricted range of the sample at the upper ability levels.

Consistent with previous research (e.g., Engelhart, 1965), there were commonalities between the phi and D analyses results for this sample. Sixteen items, or 67% of the D and 9% of the phi items, were selected from both analyses. All common items were from the CTBS (on achievement test) because the D analyses yielded no TCS (cognitive ability) NP items. Similarly, the one D analysis item found to discriminate in favor of the not-gifted students also did so on the phi analysis.

#### Cutoff Scores

Cutoff scores for the phi, D, and SIT analyses were computed in order to demarcate optimal cutting points for

differentiating the gifted from not-gifted students. Cutoff scores were established such that a maximum number of gifted and a minimum number of not-gifted students would fall above the cutoff. Therefore, the true positive and true negative findings were maximized while the false positives and false negatives were minimized. This procedure required a "value judgment" by the rater as to the amount of error which would be acceptable. As the cutoff is lowered to permit more gifted students to exceed it, increasing numbers of not-gifted students would also exceed it, thus increasing the possibility of false negatives. To assist with this problem, multiple cutoff scores were calculated so that the most desirable cutoffs could be determined. Cutoff scores were chosen according to fractions of standard deviation units for the total score distributions on the SIT, the new test created by selecting items with phi, and the new test created by selecting items with D. The mean SIT IQ for the Phase I sample was 137.30 and the standard deviation was 14.12. This mean IQ was approximately 37 points higher than that of the general population. It reflects the above average general intelligence of the research sample. In Phase II, these standard deviation cutoff scores calculated as standard deviations from the Phase I sample, were applied to the cross validation sample to determine the relative accuracy of gifted classifications using the NP-phi, NP-D, and SIT.



## Phase II--Cross Validation

### Application of Cutoff Scores

The first step in testing the NP items obtained in Phase I was to calculate the numbers of NP items scored correctly for the Phase II sample and then to apply appropriate cutoff points to those scores. The Phase II subsample consisted of 61 students. Proportions of correct classifications among the NP-phi, NP-D and SIT items were then analyzed to determine the test and cutoff that most accurately differentiated the gifted and not-gifted students.

Tables 4-3, 4-4, and 4-5 show the numbers of examinees classified into prediction categories at various cutoff points where the examinee's true status was defined as classification of gifted or not-gifted according to the WISC-R. For all three classification procedures (NP-phi, NP-D, and SIT) little variability existed between prediction rates at some of the cutoffs. For example, classifications at the +1.0 and -1.0 differed only slightly from those at the +.66 and -.66 cutoffs respectively. A similar result occurred for the +.50 cutoff compared to the +.30 cutoff and for the -.50 cutoff compared to the -.30 cutoff.

Table 4-3. Phase II: Number of Examinees in Each Prediction Category by Cutoff/Phi

			True Positive	True Negative	True Total	False Negative	False Positive	False Total
Cutoff								
-1.00	SD	30		5	35	2	24	26
- .66	SD	30		5	35	2	24	26
- .50	SD	30		5	35	2	24	26
- .33	SD	30		8	38	2	21	23
0.00	SD	29		12	41	3	17	20
+ .33	SD	23		17	40	9	12	21
+ .50	SD	18		19	37	11	10	24
+ .66	SD	8		24	32	24	5	29
+1.00	SD	3		29	32	29	0	29

Total gifted = 32

Total not-gifted = 29

Table 4-4. Phase II: Number of Examinees in Each Prediction Category by Cutoff/Index of Discrimination

			True Positive	True Negative	True Total	False Negative	False Positive	False Total
Cutoff								
-1.00	SD	31		5	36	1	24	25
- .66	SD	31		9	40	1	20	21
- .50	SD	31		9	40	1	20	21
- .33	SD	31		10	41	1	19	20
0.00	SD	26		15	41	6	14	20
+ .33	SD	26		15	41	6	14	20
+ .50	SD	22		19	41	10	10	20
+ .66	SD	22		19	41	10	10	20
+1.00	SD	15		24	39	17	5	22

Table 4-5. Phase II: Number of Examinees in Each Prediction Category Cutoff/SIT

Cut-off		True Positive	True Negative	True Total	False Negative	False Positive	False Total
-1.0	SD	32	6	38	1	22	23
- .66	SD	32	7	39	1	21	22
- .50	SD	31	7	38	2	21	23
- .33	SD	28	9	37	5	19	24
.00	SD	22	18	40	11	10	21
+ .33	SD	14	22	36	19	6	25
+ .50	SD	12	24	36	20	5	25
+ .66	SD	9	26	35	23	2	25
+1.0	SD	5	29	34	27	0	27

Table 4-6 was constructed to consolidate and clarify the data on predictive accuracy, in that table, proportions of correct and incorrect predictions are again expressed as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Deleted were the cutoffs -1.0 SD, -.50 SD, +.50 SD, and -1.0 SD. Inspection of Table 4.6 reveals that all three methods of screening gifted students were approximately equivalent. That is, all three found the greatest proportions of TP predictions in .492 through .525 of the sample. Lowest TN predictions occurred in .082 through .148 of the sample.

After the prediction rates were established for the three tests and the prediction values were expressed as proportions of the sample, analyses were conducted to determine the proportions, if any, that were most accurate. Coefficient Kappa analyses were calculated to assess the relative gifted prediction accuracy of the NP-phi, NP-D, and SIT.

#### Kappa Comparisons

In results shown in Table 4-7, P values correspond to percentages of correct classifications (both true positives and true negatives). However, two equal classification percentages (e.g., as observed for phi and D at 0.0 cutoff) may not have equal corresponding K-values because the Kappa equation adjusts for chance correct predictions.

Table 4-6. Phase II: Proportions of Correct and Incorrect Predictions for NP ( $\phi$ ), NP (D), and SIT at Two Score Cutoffs

z-Cutoff	True Positive	True Negative	True Total	False Negative	False Positive	False Total
$\phi$						
- .66 SD	.492	.082	.574	.033	.393	.426
- .33 SD	.492	.131	.623	.033	.344	.377
.00 SD	.475	.197	.672	.049	.229	.328
+ .33 SD	.377	.279	.656	.164	.180	.344
+ .66 SD	.131	.393	.525	.393	.082	.475
D						
- .66 SD	.508	.148	.656	.016	.328	.344
- .33 SD	.508	.164	.672	.016	.311	.328
.00 SD	.426	.246	.672	.098	.230	.328
+ .33 SD	.426	.246	.672	.098	.230	.328
+ .66 SD	.361	.311	.672	.164	.164	.328
SIT						
- .66 SD	.525	.115	.639	.016	.344	.361
- .33 SD	.459	.148	.607	.082	.311	.393
.00 SD	.361	.295	.656	.180	.164	.344
+ .33 SD	.270	.360	.590	.311	.082	.410
+ .66 SD	.164	.426	.590	.361	.049	.410

Note: All values rounded to nearest .10 percentile.

Table 4-7. Percentage of Correct Classifications  
and Kappa Values for NP (phi and D)  
and SIT at Various Cutoff Scores

Cutoff	NP		SIT-P*	NP		SIT-k <sup>1</sup>
	Phi-P*	D-P*		Phi-k <sup>1</sup>	D-k <sup>1</sup>	
- .66 SD	57.4	65.6	63.9	.114	.289	.234
- .33 SD	62.3	67.2	60.7	.220	.324	.178
.00 SD	67.2	67.2	65.6	.328	.334	.309
+ .33 SD	65.6	67.2	59.0	.306	.320	.258
+ .66 SD	52.5	67.2	59.0	.076	.342	.204

\*P = percentage

<sup>1</sup>k = Kappa value



An example of the Kappa calculation of percentages of correct classifications for NP-phi at the -.66 cutoff is illustrated using four-fold tables (Figure 4-1). Values in the equation, along with percentages of examinee classifications at other cutoffs, are located in Table 4-6.

$$K = \frac{P - P_C}{1 - P_C}$$

$$P = .492 + .082 = .574$$

$$\begin{aligned} P_C &= (.885)(.525) + (.475)(.115) \\ &= .4646 + .0546 \\ &= .5192 \end{aligned}$$

Results of the Kappa analyses revealed a consistent pattern of greater values for the NP-D than for the NP-phi or the SIT at all cutoffs, suggesting the NP-D may be the more accurate predictor of correct classification. The greatest K value for NP-D occurs at the +.66 where the K value of .342 indicates that there is a 34% improvement in prediction accuracy over that expected by chance. The greatest discrepancy between K values for NP-D and the SIT appears at the -.33 cutoff where  $K = .324$  for NP-D but only .178 for the SIT.

		IQ		
		Class Gifted	Class Not-Gifted	
phi	Class Gifted	TP .492	FP .393	.885
	Class Not-Gifted	FN .033	TN .082	.115
		.525	.475	

$$K = \frac{.574 - .519}{1 - .519}$$

$$= \frac{.055}{.481} = .114$$

Figure 4-1. Computation Example of Kappa analysis for True Phi Classifications at  $-.66$  SD Cutoff

### Summary of Results

In Phase I of this study 56 items from the CTBS/TCS were identified using phi coefficients that significantly differentiated gifted from not-gifted students. The index of discrimination analyses yielded 24 items that successfully discriminated gifted from not-gifted. A large proportion of significant D items were common to the phi analysis results.

In Phase II, total scores for each subset of items were computed using a cross validation sample of examinees. The accuracy of classifications resulting from each of the two new tests and the SIT were contrasted at five different cutoff scores. Kappa analyses revealed higher correct prediction values for the NP-D than for either the NP-phi and the SIT at all cutoff levels.

## CHAPTER V DISCUSSION

### Research Questions

In order to answer the first research question, this discussion focuses on issues related to NP item validity from Phase I. Phase II NP items were selected from the CTBS and TCS using correlational analyses. To answer the remaining two research questions, the predictive accuracies of the NP-phi, NP-D, and SIT are discussed in terms of cutoff scores.

The following research questions were addressed in this study.

1. Can an accurate predictor of gifted IQ classification on the WISC-R/S-B be derived from an instrument composed of items on the CTBS and TCS in a situation in which giftedness is viewed as a dichotomous trait variable?
2. Is the NP more accurate than the SIT in classifying gifted and not-gifted seventh graders?
3. At what cutoff point(s) is the NP more accurate than the SIT in classifying gifted and not-gifted seventh graders?

### Phase I

In Phase I of the study it was suggested that some items derived from the CTBS/TCS were valid for distinguishing between gifted and not-gifted seventh graders. Evidence for this assertion included (a) the relatively large number of NP items that are common to both the phi and D analyses, (b) the large number of items found to discriminate between the two groups, and (c) the agreement between CTBS and IRT location parameter estimates and NP item statistics.

### Common Items

The index of discrimination (D) analyses yielded fewer discriminating items than did the phi analyses. However, the validity of the D analyses is supported by the overlap of items between the phi and D analyses. Of the 14 items retained for the NP-D, 11 items were also included in the phi NP scale. Overlap between phi and D item discrimination for middle difficulty items was reported by Engelhart (1965).

Support for the validity of the NP also was suggested by the relatively large number of significant items, particularly on the phi analyses. In addition to the 56 significant phi items initially computed at  $\leq .05$ , another 29 were significant at the .10 level. Clearly an even larger number of significant items would have resulted had the

sample size of 117 remained constant. However, the sample size diminished below 75 on 99 items. Of these items, 20 had phi values that would have been significant at  $<.05$  had the N been as high as 113.

#### IRT Parameters

A strong case for the NP item validity is made upon investigation of item response theory item location parameters for the CTBS items. An investigation of the response patterns of subjects on both item sets, in comparison with the response pattern on these items by the original CTBS sample, suggest commonality in responding.

IRT item location parameter. Item response theory was utilized by the CTBS/TCS constructors for item selection. Item characteristic curves were plotted using a three parameter logistic model involving (a) item discrimination, (b) item location (or difficulty), and (c) a "guessing" factor. A location parameter describes an item's difficulty in terms of the student's ability level (or latent trait). An item discriminates best for a student whose ability level is near the item's location parameter (or difficulty level). An item with a high location parameter serves its function only for students of high ability since low ability students would not be expected to answer these items correctly.

Comparing response patterns. An assumption of item response theory is the invariance of item parameters

(Anastasi, 1982; Baker 1984). That is, given certain conditions, item parameters should be uniform among different populations because individual items are assumed to measure the same trait in different populations. In this case, the validity of the NP items would be supported if NP items were common to CTBS items with high location parameters. In fact, 41 of the original 65 NP items had CTBS location parameters above the mean for the norming sample (CTBS Technical Report, 1983). The proportion of NP-D items to high location parameter items was even greater (16 of 24, 67%).

In summary, the data support an affirmative answer to the first research question. There exists a subset of items that discriminated gifted and not-gifted seventh graders who were all academic high achievers. In Phase II, cross validation with a smaller but otherwise equivalent sample was conducted to support these findings.

## Phase II

Cross validation of Phase I findings was conducted in Phase II by assessing the accuracy of NP items as compared to the SIT in classifying the gifted and not-gifted students and the cutoff score at which each procedure showed greatest accuracy. These goals were accomplished by analyzing proportions of correct and incorrect classifications using coefficient Kappa analyses.



An examination of total true classifications (i.e., correct classifications of gifted and not-gifted students) revealed greatest prediction accuracy at the 0.0 SD cutoff. At this cutoff, NP-D and NP-phi values were equal (67.2) and slightly superior to the SIT (65.6). When percentage values were corrected for chance occurrences by the Kappa analyses, NP-D was slightly more accurate at the +.66 SD than at 0.0 SD, and NP-D items were superior to NP-phi and the SIT. Furthermore, at the +.66 cutoff, the Kappa value for the NP-D is .342 contrasted to much lower values for NP-phi (.076) and SIT (.204).

In answering the second research question, data consistently supported the NP as more accurate than the SIT in classifying gifted and not-gifted students. Among the two NP tests, NP-D was generally superior to the NP-phi. The NP-D was clearly the better measure of true positive and negative classifications based on Kappa analyses.

There is some ambiguity regarding the overall most accurate cutoff for classifying students, rendering the last research question more difficult to answer. There is empirical support for the NP-D at both the -.33 and +.66 cutoffs as most accurate for classifying gifted and not-gifted seventh graders in the sample. Kappa analyses support that finding in that the discrepancy between total

true classifications for the NP-D and the SIT was greatest at the  $-.33$  cutoff. Data supporting an NP-D cutoff at  $+.66$  may be less clear. Kappa analyses indicate that the number of total true classifications at  $+.66$  is superior to the number at any other cutoff. The relative strength of the NP-D in total true classifications was due to its much greater accuracy in predicting true positives. The SIT was more accurate in classifying true negatives. In choosing a cutoff score for gifted IQ screening purposes, specific classification priorities and goals must be taken into account. For example, if the primary goal is to maximize total true classifications, the  $+.66$  cutoff may be desirable.

### Conclusions, Implications, and Limitations

This study yields conclusions, implications, and limitations germane to both pragmatic and theoretical issues. In the following sections, the issues of sampling, generalizability, item validity, and cutoff scores are discussed individually.

#### Sampling

The success of the NP in classifying two groups of students, who in some respects appeared indistinguishable, has some inherent implications. The gifted and not-gifted children were similar in terms of achievement, IQ (in many

cases), and teacher perceptions of them as gifted. These similarities suggest that the analysis results are robust because the two groups were accurately differentiated in spite of their likenesses.

The specific trait of the research sample as being high academic achievers distinguishes this study from many others and this distinction is essential to the utility of the NP. In most of the related literature reviewed in Chapter II, research samples were not as restricted in range of IQ, rendering correct classification of gifted and not-gifted more likely in those. In other studies, many samples included students randomly selected from general populations. Those students, excluded in this study by preselection, were readily screened as not-gifted in other studies, allowing for artificially inflated accuracy predictions.

A pertinent limitation to this study discussed in Chapter III was the sampling procedure. Specifically, some students were deleted from formal IQ testing, and therefore, excluded from the sample because they scored below the SIT screening cutoff of 132. However, one-sixth (i.e., 11 of 61) Phase II students had SIT scores below 132, with some scores falling in the average and below average range. This occurrence raises a question about the criterion used for disqualifying students from IQ

testing based on their SIT scores. It is unclear how students with SIT scores in the 80s were not disqualified while others were. It seems that screening procedures were not followed consistently at the school level. Fortunately, even though there were relatively few students with SIT IQs below 132, there were enough for data analyses, and the range of scores was wide. In this case, sampling error probably was not a confounding influence on results because disqualification of the aforementioned students seems to have occurred in a random, unsystematic manner.

#### Generalizability

In large part, outcomes regarding screening accuracy, item validity, and cutoff scores may not be generalizable to other populations at this time. However, the success of the research procedure (NP) is considered meaningful in as much as the accuracy of the NP will be tested continually in its practical application on potentially gifted students. In this way, the generalizability of screening accuracy, item validity, and cutoff scores will be validated on other populations.

In Chapter I the definition of intellectual giftedness for this study was briefly discussed. This pragmatic definition focuses only on IQ test scores that fall above

a particular cutoff score. Other, more theoretical definitions, may involve other qualitative differences. Results of this study may not be generalizable to students who are designated gifted using criteria other than their IQs falling above the 96th percentile.

### Screening Accuracy

Generalizability of research results would be enhanced by analyzing group test data for other restricted populations to determine if subgroups may be successfully discriminated using test items. Two preliminary steps in such a process would be to (a) analyze test results of other populations of potentially gifted students, such as elementary or high school, on the CTBS/TCS, and (b) determine if other group achievement/cognitive ability tests possess items that accurately classify potentially gifted students. Further, researchers may seek to generalize results on more divergent restricted populations. For example, it may be useful to screen for mild mental retardation among remedial students or for learning disabilities among children with discrepant report card grades.

### Item Validity

A somewhat unexpected result of this research was that the CTBS items (which supposedly measure school achievement) contributed much more to the NP than did the TCS items (which are purported to measure intelligence).

Indeed, none of the NP-D items were obtained from the TCS. The NP-phi test, which initially included eight TCS items, was found to be generally less accurate in classifying students than the NP-D. Essentially, the achievement items had greater criterion related validity than did the cognitive ability items for Full Scale IQ.

Ostensibly, the superior criterion validity of the CTBS items is contrary to expectation, however a closer examination suggests that this pattern may support rather than refute the generalizability of research findings. Examination of CTBS, TCS, and WISC-R subtests reveals that the CTBS probably has more in common with the WISC-R than does the TCS. For example, CTBS subtests such as Mathematics, Science, Social Studies, and Vocabulary have what appear to be direct correlates on the WISC-R. The relationship between TCS and WISC-R subtests appears more obscure. This assertion is supported by Wurster (1985) who compared overlap between the TCS and WISC-R with the SIT and the WISC-R. She found that 87.8% of the SIT items measured the same skills as the WISC-R Verbal Scale, however, "no items from the TCS appeared to measure any of the skills that are measured by the 11 WISC-R subtests" (p. 24).

The relationship of commonality between IQ test and achievement test performance for this sample seems related

to two factors. First, it is likely that for high achieving students, "superior" intelligence is heavily loaded in the verbal reasoning domains that are tapped to a greater extent by the CTBS than by the TCS. The TCS, to a greater degree than the CTBS, measures non-verbal skills or abilities that may be less represented by high achieving students' strengths on the WISC-R or Stanford-Binet. Future researchers might investigate this hypothesis by examining the relationship between achievement and verbal vs. non-verbal intelligence (as represented by these or similar tests) for high achieving or gifted students.

A second explanation for the disproportionate representation of achievement test items on the NP concerns the theoretical rationale for this research. That is, the content of achievement tests (e.g., CTBS) and IQ tests (e.g., WISC-R) are often operationally indistinguishable. This premise has been supported empirically by Anastasi (1982), among many others (see Chapter II).

### Cutoffs

Conclusions regarding the most desirable cutoff for the NP (NP-D at  $-.33$  or  $+.66$ ) were discussed earlier in the chapter. There are some other conclusions that may be drawn regarding cutoff scores for the SIT with this population. The mean SIT score for the Phase II sample was 137. This score is somewhat above what would be expected yet consistent with research findings by Karnes and Brown



(1979) and Rust and Lose (1980) who found that the SIT tends to overestimate high IQs on full length tests. In their studies, those researchers recommended setting SIT cutoff scores for gifted classification higher than two standard deviations above the mean to offset this tendency to overestimate the IQs of brighter students. The current findings support those recommendations because the SIT was most accurate in classifying true positive and false negative findings at over three standard deviations above the mean (SIT IQ = 149).

#### Summation

In his later years Edwin R. Guthrie (1959) suggested that research has no inherent value, but rather that its value was gained from its practical applications. It is in the spirit of that philosophy that this research may be fully appreciated. Relatively accurate in its predictive power and efficient in its method, the NP analysis will be easily replicated on new student samples and test formats. Thus, the NP may be best viewed as a procedure adaptable to varying, yet specific, needs.

## REFERENCES

- Ahmann, J. S. (1972). The comprehensive test of basic skills (a review). In O. K. Buros (Ed.), The seventh mental measurement yearbook (pp. 614-615). Highland Park, NJ: The Gryphon Press.
- American Psychological Association (1985). Standards for educational and psychological tests. Washington, DC: American Psychological Association.
- Anastasi, A. (1976). Psychological testing. New York: MacMillan Publishing.
- Apple, D. (1983, November). Screening gifted children: A comparison of the SIT and WISC-R. Paper presented at the annual meeting of the Alabama Association of School Psychologists, Guntersville, AL.
- Barklay, J. E., Phillips, G., & Jones, T. (1983). Developing a predictive index for giftedness. Measurement and Evaluation in Guidance, 16, 25-35.
- Barrington, B. L. (1979). In the name of education. In N. Colangelo and R. T. Zarram (Eds.), New voices in counseling the gifted (pp. 65-70). Dubuque, Iowa: Kendall/Hunt.
- Bersoff, D. N. (1971). Short forms of individual intelligence tests for children: Review and critique. Journal of School Psychology, 9, 310-320.
- Birch, J. W. (1955). The utility of short forms of the Stanford Binet tests of intelligence with mentally retarded children. American Journal of Mental Deficiency, 59, 462-484.
- Birch, J. W. (1984). Is any identification procedure necessary? Gifted Child Quarterly, 28, 157-161.
- Bloom, B. S. (1956). Taxonomy of educational objectives handbook I: The cognitive domain. New York: David McKay Co.

- Blosser, G. H. (1963). Group intelligence tests as screening devices in locating gifted and superior students in ninth grade. Exceptional Children, 29, 282-286.
- Braden, J. P. (1985). A modest proposal: Using probabilities of special-education eligibility instead of cutting scores. Unpublished manuscript.
- Brock, H. (1982). Factor structure of intellectual and achievement measures for learning disabled children. Psychology in the Schools, 19, 297-304.
- Brooks, C. R. (1977). WISC, WISC-R, S-B L-M, WRAT: Relationships and trends among children ages six to ten referred for psychological evaluation. Psychology in the Schools, 14, 30-33.
- Bryan, J. R., & Bryan, T. H. (1975). Understanding learning disabilities. Sherman Oaks, CA: Alfred Publishing Co.
- Burket, G. R. (1974). Empirical criteria for distinguishing and validating aptitude and achievement measures. In D. A. Green (Ed.), The aptitude-achievement distinction (pp. 35-51). Monterey, CA: CTB/McGraw-Hill.
- Carleton, F. O., & Stacey, C. L. (1954). Evaluation of selected short forms of the Wechsler Intelligence Scale for Children. Journal of Clinical Psychology, 10, 258-261.
- Carroll, J. B. (1966). Factors of verbal achievement. In A. Anastasi (Ed.), Testing problems in perspective (pp. 406-413). Washington, D.C. American Council on Education.
- Chambers, J. A. (1960). Preliminary screening methods in the identification of intellectually superior children. Exceptional Children, 26, 145-150.
- Chambers, J. A., Barron, F., & Sprecher, J. W. (1980). Identifying gifted Mexican-American students. Gifted Child Quarterly, 24, 123-128.
- Clarizio, H. F., & Meherens, W. A. (1985). Psychometric limitations of Guilford's structure-of-intellect model for identification and programming of the gifted. Gifted Child Quarterly, 29, 113-119.

- Cohen, J. A. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.
- Crofoot, M. J., & Bennett, T. S. (1980). A comparison of three screening tests and the WISC-R in special education evaluations. Psychology in the Schools, 17, 474-478.
- CTB/McGraw-Hill. (1984). The Comprehensive Test of Basic Skills: Technical report. Monterey, CA: CTB/McGraw-Hill.
- CTB/McGraw-Hill. (1984). The Test of Cognitive Skills: Technical report. Monterey, CA: CTB/McGraw-Hill.
- Dean, R. S. (1977). Canonical analysis of a jangle fallacy. Multivariate Experimental Clinical Research, 3, 17-20.
- Dean, R. S. (1982). Intelligence-achievement discrepancies in diagnosing pediatric learning disabilities. Clinical Neuropsychology, 4, 58-62.
- DeFilippis, N. A., & Fulmar, K. (1980). Effects at age and IQ level on the validity of one short intelligence test used for screening purposes. Educational and Psychological Measurement, 40, 543-545.
- Dirkes, M. A. (1981). Only the gifted can do. Educational Horizons, 59, 138-143.
- Dirks, J., Wessels, K., Quaforth, J., & Quenon, B. (1980). Can short form WISC-R tests identify children with high Full Scale IQ? Psychology in the Schools, 17, 40-46.
- Dunn, L. M., & Markwardt, F. C. (1970). Peabody Individual Achievement Test. Circle Pines, MN: American Guidance Service.
- Ebel, R. L. (1965). Measuring educational achievement. Englewood Cliffs, NJ: Prentice-Hall.
- Elman, L., Blixt, S., & Sawicki, R. (1981). The development of cutoff scores on a WISC-R in the multidimensional assessment of gifted children. Psychology in the Schools, 43, 426, 428.

- Enburg, R., Rowley, V. N., & Stone, B. (1961). Short forms of the WISC for use with emotionally disturbed children. Journal of Clinical Psychology, 17, 280.
- Engelhart, M. D. (1965). A comparison of several item discrimination indices. Journal of Educational Measurement, 2, 69-76.
- Fell, L., & Fell, S. S. (1982). Effectiveness of WISC-R short-forms in screening gifted children. Psychological Reports, 51, 1017-1018.
- Findley, C. J., & Thompson, J. M. (1958). An abbreviated Wechsler Intelligence Scale for Children for use with educable mentally retarded. American Journal of Mental Deficiency, 63, 473-480.
- Friedes, D. (1978). Review of the WISC-R. In O.K. Buros (Ed.), The eighth mental measurement yearbook (pp. 414-422). Highland Park, NJ: The Gryphon Press.
- Gensley, J. (1979). Parent perspectives on the curiosity quotient. Gifted Child Quarterly, 23, 723-724.
- Gronlund, N. E. (1976). Measurement and evaluation in teaching. New York: McMillan Publishing Co.
- Grossman, F. M., & Galvin, G. A. (1987). Clinically and theoretically derived WISC-R subtest regroupings: Predicting academic achievement in a referral population. Psychology in the Schools, 24, 105-108.
- Grossman, F. M., & Johnson, K. M. (1982). WISC-R factor scores as predictors of WRAT performance: A multivariate analysis. Psychology in the Schools, 19, 465-468.
- Grossman, F. M., & Johnson, K. M. (1983). Validity of the Slosson and Otis-Lennon in predicting achievement of gifted students. Educational and Psychological Measurement, 43, 617-622.
- Guilford, J. P. (1975). Varieties of creative giftedness, their measurement and development. Gifted Child Quarterly, 19, 107-121.
- Guthrie, E. R. (1959). Association by contiguity. In S. Koch (Ed.), Psychology, a study of science (pp. 158-195). New York: McGraw-Hill.

- Hale, R. L. (1978). The WISC-R as a predictor of WRAT performance. Psychology in the Schools, 15, 172-175.
- Hammill, D., & McNutt, G. (1981). The correlates of reading. Austin, TX: Pro-Ed.
- Harrington, R. G. (1982). Standardized testing may be hazardous to the educational progress of intellectually gifted children. Education, 103, 112-117.
- Hartlage, L. C., & Steele, C. T. (1977). WISC and WISC-R correlates of academic achievement. Psychology in the Schools, 14, 15-18.
- Hirsch, F. J., & Hirsch, S. J. (1980). The Quick Test as a screening device for gifted students. Psychology in the Schools, 17, 37-46.
- Horn, J. L. (1970). Factor analysis with variables of different metric. Educational and Psychological Measurement, 29, 753-762.
- Hurrocks, J. E. (1964). Assessment of behavior, the methodology and content of psychological measurement. Columbus, OH: Charles E. Merrill Publishing.
- Jenkins-Friedman, R. (1982). Myth: Cosmetic use of multiple situation criteria. Gifted Child Quarterly, 26, 24-26.
- Jensen, A. R. (1980). Bias in mental testing. New York: McMillan.
- Jensen, A. R. (1981). Straight talk about mental tests. New York: Free Press.
- Joesting, J., & Joesting, R. (1971). The Quick Test as a screening device in a welfare setting. Psychological Reports, 29, 1289-1290.
- Karnes, F. A., & Brown, K. E. (1979). Comparison of the SIT with the WISC-R for gifted students. Psychology in the Schools, 16, 478-482.
- Karnes, F. A., & Brown, K. E. (1980). Factor analysis of the WISC-R for the gifted. Journal of Educational Psychology, 72, 197-199.
- Karnes, F. A., & Brown, K. E. (1981). A short form of the WISC-R for gifted students. Psychology in the Schools, 18, 169-173.



- Karnes, F. A., Edwards, R. P., & McCallum, R. S. (1986). Normative achievement assessment of gifted children: Comparing the K-ABC, WRAT, and CAT. Psychology in the Schools, 23, 346-352.
- Kaufman, A. S. (1975). Factor analysis of the WISC-R at eleven age levels between 6 1/2 and 16 1/2 years. Journal of Consulting and Clinical Psychology, 43, 135-147.
- Kaufman, A. S. (1979). Intelligence testing with the WISC-R. New York: John Wiley and Sons.
- Kendall, P. C., & Little, V. L. (1977). Correspondence of brief intelligence measures to the Wechsler scales of delinquents. Journal of Consulting and Clinical Psychology, 45, 660-666.
- Killian, J. B., & Hughes, C. D. (1978). A comparison of short forms of the Wechsler Intelligence Scale for Children-Revised in the screening of gifted referrals. Gifted Child Quarterly, 22, 111-115.
- Kolloff, P. B., & Feldhusen, J. F. (1984). The effects of enrichment on self concept and creative thinking. Gifted Child Quarterly, 28, 53-57.
- Kramer, J. J., Markley, R. P., Shanks, K., & Ryabik, J. E. (1983). The seductive nature of WISC-R short forms: An analysis with gifted referrals. Psychology in the Schools, 20, 137-141.
- Lawrence, D., & Anderson, H. N. (1979). A comparison of the Slosson Intelligence Test and the WISC-R with elementary school children. Psychology in the Schools, 16, 361-364.
- Lennon, R. T. (1978, March). Perspective on intelligence testing. Address presented at the National Council on Measurement in Education, Toronto.
- Lorr, M., & Meister, R. K. (1942). The optimum use of test data. Educational and Psychological Measure-

- Male, R. A., & Perrone, P. (1979). Identifying talent and giftedness. Roeper Review, 2, 5-11.
- Mallinson, G. G. (1963). An analysis of the factors related to the motivation and achievement of students in science courses in the junior and senior high, final report (Report No. CRP-503). Kalamazoo, MI: Western Michigan University School of Graduate Studies. (ERIC Document Reproduction Service No. ED 002889)
- Martin, J. D., & Kidwell, J. C. (1977). Intercorrelations of the Wechsler Intelligence Scale for Children-Revised, the Slosson Intelligence Test, and the National Educational Developmental Test. Educational and Psychological Measurement, 37, 1117-1120.
- Martin, J. D., & Rudolph, L. (1972). Correlates of the Wechsler Adult Intelligence Scale, the Slosson Intelligence Test, ACT scores and grade point averages. Educational and Psychological Measurement, 32, 459-462.
- Mayfield, B. (1979). Teacher perception of creativity, intelligence and achievement. Gifted Child Quarterly, 23, 812-817.
- McNemar, Q. (1962). Psychological statistics (3rd Ed.). New York: John Wiley.
- Meister, R. K., & Kurko, V. K. (1951). An evaluation of a short administration of the revised Stanford-Binet Intelligence Examination. Educational and Psychological Measurement, 11, 489-493.
- Mercer, J. R. (1979). In defense of racially and culturally non-discriminatory assessment. School Psychology Digest, 8, 89-115.
- Mize, J. M., Smith, J. W., & Callaway, B. (1979). Comparison of reading disabled childrens scores on the WISC-R, Peabody Picture Vocabulary Test and Slosson Intelligence Test. Psychology in the Schools, 16, 356-358.
- Nichols, J. E. (1962). Brief forms of the Wechsler Intelligence scales for Research. Journal of Clinical Psychology, 18, 167.



- Nicholson, C. L. (1977). Correlations between the Quick Test and the Wechsler Intelligence Scale for Children-Revised. Psychological Reports, 40, 523-526.
- Olsen, A. V., & Rosen, C. L. (1971, February). Exploration of the structure of selected reading readiness tests. Paper presented at the meeting of the American Educational Research Association, New York.
- Pearce, N. (1983). A comparison of the WISC-R, Raven's Standard Progressive Matrices, and Meeker's S0I-Screening Form for gifted. Gifted Child Quarterly, 27, 13-18.
- Pedriana, A. J., & Bracken, B. A. (1982). Performance of gifted children on the PPVT and PPVT-R. Psychology in the Schools, 15, 183-185.
- Pegnato, C. W., & Birch, J. W. (1959). Locating gifted children in junior high schools--A comparison of methods. Exceptional Children, 25, 300-304.
- Petty, M. F., & Field, C. J. (1980). Fluctuations in mental test scores. Educational Research, 22, 198-202.
- Reschley, D. J., & Reschley, J. E. (1979). Validity of WISC-R factor scores in predicting achievement and attention for four socio-cultural groups. Journal of School Psychology, 17, 355-361.
- Ricca, J. (1984). Learning styles and preferred instructional strategies of gifted students. Gifted Child Quarterly, 12, 121-126.
- Ritter, D., Duffy, J., & Fischman, R. (1973). Comparability of Slosson and S-B estimates of intelligence. Journal of School Psychology, 3, 224-227.
- Roach, P. A. (1979). The effects of conceptual style preference, related cognitive variables and sex on achievement in mathematics. British Journal of Educational Psychology, 49, 79-82.
- Rust, J. D., & Lose, B. D. (1980). Screening for giftedness with the Slosson and the Scale for Rating Behavioral Characteristics of Superior Students. Psychology in the Schools, 17, 446-451.

- Ryan, G. T. (1979). The influence of readability on critical reading comprehension of secondary social studies students. Dissertation Abstracts International, 39, 6684. (University Microfilms No. 79-11,017)
- Ryan, G. T. (1982). An analysis of motivational factors of gifted and non-gifted learners. Paper presented at the annual meeting of the American Educational Research Association, New York.
- Salvia, J., & Ysseldyke, J. E. (1978). Assessment in special and remedial education. Boston: Houghton Mifflin Co.
- Schena, R. A. (1963). A search for talented pupils. Journal of Experimental Education, 32, 27-41.
- Schnell, R. E. (1982). Is there a correlation between intelligence and reading achievement for students in grades three through eight who were referred to the gifted program in Hendry County Florida. Unpublished manuscript.
- Schwarting, F. G., & Schwarting, K. R. (1977). The relationship of the WISC-R and WRAT: A study based upon a selected population. Psychology in the Schools, 14, 431-433.
- Sheldon, W. D., & Manolakes, G. (1954). A comparison of the Stanford-Binet revised form L, and the California Test of Mental Maturity (S-Form). Journal of Educational Psychology, 45, 499-504.
- Silverstein, A. B. (1970). Reappraisal of the validity of a short form of Wechsler's Scales. Psychological Reports, 26, 559-561.
- Simpson, W. H., & Bridges, C. C. (1959). A short form of the Wechsler Intelligence Scale for Children. Journal of Clinical Psychology, 15, 414.
- Slosson, R. J. (1961). Slosson Intelligence Test for Children and Adults. East Aurora, NY: Slosson Educational Publications.
- Slosson, R. J., & Jensen, J. A. (1982). Slosson Intelligence Test (SIT) norms tables application and development. East Aurora, NY: Slosson Educational Publications.

- Stedman, J. N., Lawlis, G. F., Cortner, R. H., & Achtenberg, D. (1978). Relationships between WISC-R factors, Wide Range Achievement Test scores, and visual motor maturation in children referred for psychological evaluation. Journal of Consulting and Clinical Psychology, 46, 869-872.
- Stenson, C. M. (1982). Note on concurrent validity of structure of intellect gifted screener with Wechsler Intelligence Scale for Children-Revised. Psychological Reports, 50, 552.
- Sternberg, R. J. (1982). Lies we live by: Misconceptions of test in identifying the gifted. Gifted Child Quarterly, 26, 157-161.
- Stewart, K. D., & Jones, E. C. (1976). Validity of the Slosson Intelligence Test: A ten year review. Psychology in the Schools, 13, 372-380.
- Stewart, D. W., & Morris, L. (1977). Intelligence and academic achievement in a clinical adolescent population. Psychology in the Schools, 14, 513-518.
- Terman, L. M., & Merrill, M. A. (1973). Stanford-Binet Intelligence Scale. Boston: Houghton Mifflin.
- Thompson, J., & Findley, C. J. (1962). The validation of an abbreviated Wechsler Intelligence Scale for Children for use with the educable mentally retarded. Educational and Psychological Measurement, 22, 539-542.
- Tuttle, F. B., & Becker, L. A. (1980). Characteristics and identification of gifted and talented students. Washington: National Educational Association.
- Undheim, J. O. (1976). Ability structure in 10-11-year-old children and the theory of fluid and crystallized intelligence. Journal of Educational Psychology, 4, 411-423.
- Vandiver, P. C., & Vandiver, S. S. (1979). A "nonbiased assessment" of intelligence testing. The Educational Forum, 44, 97-108.
- Vernon, P. E. (1961). The structure of human abilities. London: Methuen.

- Vernon, P. L., Adamson, G., & Vernon, D. F. (1977). The psychology and education of gifted children. Boulder, CO: Westview Press.
- Wade, D. L., Phelps, L., & Falasco, S. (1986). Use of an abbreviated version of the WISC-R with learning disabled children. Psychology in the Schools, 23, 353-356.
- Washington, E. D., Engelmann, S., & Bereiter, C. (1969). Achievement components of Stanford-Binet performance. (ERIC Document Reproduction Service No. ED 056771)
- Wechsler, D. (1950). Cognitive, conative, and non-intellective intelligence. American Psychologist, 5, 78-83.
- Wechsler, D. (1958). The measurement and appraisal of adult intelligence. Baltimore: Williams and Wilkins.
- Wechsler, D. (1974). Wechsler Intelligence Scale for Children-Revised. New York: The Psychological Corp.
- Wikoff, R. C. (1978). The WISC-R as a predictor of achievement. Psychology in the Schools, 16, 364-366.
- Wright, D. (1983). Effectiveness of the PPVT-R for screening gifted students. Psychology in the Schools, 20, 25-26.
- Wright, D., & Dappen, L. (1982). Factor analysis at the WISC-R and the WRAT with a referral population. Journal of School Psychology, 20, 306-312.
- Wright, B. W., & Sandry, M. (1962). A short form of the Wechsler Intelligence Scale for Children. Journal of Clinical Psychology, 18, 1966.
- Wurster, J. A. (1985). Gifted screening: The Slösson Intelligence Test versus the Test of Cognitive Skills. Unpublished manuscript.
- Yakowitz, J. M., & Armstrong, R. G. (1955). Validity of short forms of the Wechsler Intelligence Scale for Children (WISC). Journal of Clinical Psychology, 11, 275-277.
- Yapp, K. O. (1977) Relationships between amount of reading activity and reading achievement. Reading World, 17, 23-29.

- Yarborough, G. H., & Johnson, R. A. (1983). Identifying the gifted: A theory-practice gap. Gifted Child Quarterly, 27, 135-138.
- Yule, W., Gold, R. P., & Busch, C. (1981). WISC-R correlates of academic attainment at 16 1/2 years. British Journal of Educational Psychology, 51, 237-240.
- Zimet, S. G., Farley, G. K., & Dahlen, N. W. (1985). An abbreviated form of the WISC-R for use with emotionally disturbed children. Psychology in the Schools, 72, 19-22.

## BIOGRAPHICAL SKETCH

Randy Evan Schnell was born in Cleveland, Ohio, on January 24, 1955. Reared in Hollywood, Florida, from age seven, he attained his B.A. in psychology with a minor in education from Florida Atlantic University in 1977. Pursuing a field of graduate studies that encompassed both psychology and education, Randy attained his Ed.S. (1983) and Ph.D. (August, 1987) from the University of Florida.

Professionally, Randy has worked as a school psychologist in both urban and rural settings in Florida. He is currently working at Memphis City Schools Mental Health Center in Tennessee where he is coordinator of the Adolescent Sex Offender Program of the Memphis, Shelby County Sex Abuse Project.

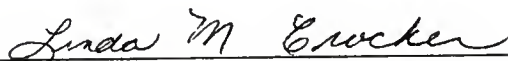


I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



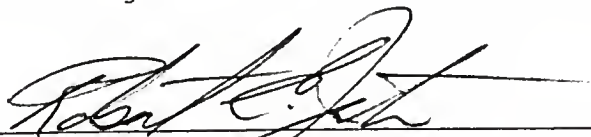
Larry C. Loesch, Chairman  
Professor of Counselor Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



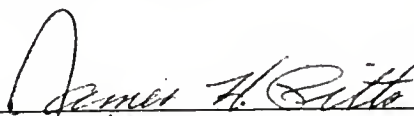
Linda M. Crocker  
Professor of Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



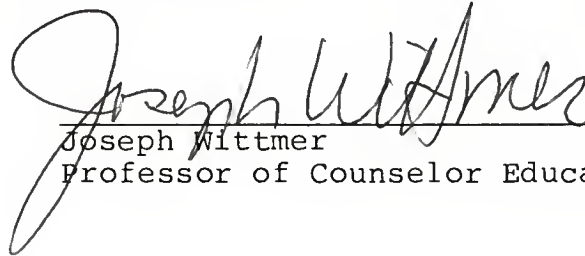
Robert E. Jester  
Professor of Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



James H. Pitts  
Assistant Professor of Counselor  
Education

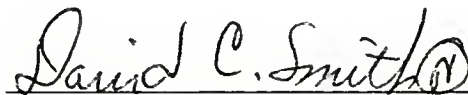
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

A handwritten signature in cursive script, reading "Joseph Wittmer", written over a horizontal line.

Joseph Wittmer  
Professor of Counselor Education

This dissertation was submitted to the Graduate Faculty of the College of Education and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 1987

A handwritten signature in cursive script, reading "David C. Smith", written over a horizontal line.

Dean, College of Education

---

Dean, Graduate School



UNIVERSITY OF FLORIDA



3 1262 08554 6736